

Computationally efficient and statistically accurate conditional independence testing with spaCRT

Ziang Niu, Jyotishka Ray Choudhury, Eugene Katsevich

March 27, 2025

Abstract

We introduce the saddlepoint approximation-based conditional randomization test (spaCRT), a novel conditional independence test that effectively balances statistical accuracy and computational efficiency, inspired by applications to single-cell CRISPR screens. Resampling-based methods like the distilled conditional randomization test (dCRT) offer statistical precision but at a high computational cost. The spaCRT leverages a saddlepoint approximation to the resampling distribution of the dCRT test statistic, achieving very similar finite-sample statistical performance with significantly reduced computational demands. We prove that the spaCRT p -value approximates the dCRT p -value with vanishing relative error, and that these two tests are asymptotically equivalent. The validity of spaCRT is assessed with modern regression techniques applied such as lasso and kernel ridge regression. Through extensive simulations and real data analysis, we demonstrate that the spaCRT controls Type-I error and maintains high power, outperforming other asymptotic and resampling-based tests. Our method is particularly well-suited for data with low signal-to-noise ratio and large scale like single-cell CRISPR screen analyses, facilitating the efficient and accurate assessment of perturbation-gene associations.

1 Introduction

Conditional independence testing is a fundamental problem in statistics. Given the joint distribution $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \sim \mathcal{L}_n$ (potentially depending on sample size n), the statistical task is to test the null hypothesis

$$H_0 : \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}. \quad (1)$$

In the language of causal inference (Imbens and Rubin, 2015), \mathbf{X} is often interpreted as a treatment variable, \mathbf{Y} as an outcome variable, and \mathbf{Z} as a set of covariates. The null hypothesis H_0 states that the treatment has no effect on the outcome, given the covariates. Suppose the independent and identically distributed observations $(X_{in}, Y_{in}, Z_{in}) \stackrel{\text{i.i.d.}}{\sim} \mathcal{L}_n$ are collected for subject $i = 1, \dots, n$. We denote these observations collectively as $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n \times d}$.

1.1 Challenges: low signal-to-noise ratio and large scale data

One particular challenge arises in modern applications involving testing conditional independence null (1), is the low signal-to-noise (SNR) ratio in the data. Such low SNR can arise from high sparsity (i.e. excessive zeros) in the observed data X and Y and also the potential heavy tail behavior of these observations. In fact, the issue of sparsity is ubiquitous in the analysis of genomic data, especially in the single-cell resolution level. Moreover, the sparsity is not the single issue involved and the large scale in such data is another barrier further precluding the application of sophisticated yet computationally intensive statistical methods. We discuss one concrete example originating from single-cell CRISPR screens analysis.

One of the biological objectives in single-cell CRISPR screens (Dixit et al., 2016; Adamson et al., 2016; Jaitin et al., 2016; Datlinger et al., 2017) is to understand *regulatory elements*, segments of DNA whose role is to control the expressions of one or more nearby genes. To address this question, single-cell CRISPR screens are designed to subject a population of cells to a large number of *CRISPR perturbations*, each of which inhibits the functioning of a specific regulatory element. Each cell receives several of these perturbations (treatment indicator $\mathbf{X} \in \{0, 1\}$), and the expression of each gene in each cell (count-based outcome $\mathbf{Y} \in \mathbb{N}$) is measured by single-cell RNA-sequencing. The statistical analysis task is to determine, for each CRISPR perturbation and gene of interest, whether cells with the perturbation have different gene expression levels compared to cells without the perturbation, subject to some collected technical covariate \mathbf{Z} such as library size and batch effect. The challenges arise in two-fold:

Sparsity in perturbations X and gene expression Y : The excessive sparsity in the treatment indicator of X , because of the pooling nature of large scale perturbations, and relatively small gene expression count Y , because of single-cell resolution and thus low coverage in RNA sequencing, can lead to a challenge of low effective sample size. Such challenge manifests itself when applying tests based on central limit theorem, in the inflated Type-I error rates (Barry et al., 2024).

Large scale in cells and tests: Sample size (number of cells) in such experiments can be as high as hundreds of thousands. Moreover, computation challenge can be even more pronounced when considering a large number of hypotheses. This is usually the practice due to the screening nature of such experiments and thus excludes the possibility of direct application of method like resampling-based test procedures despite they are able to produce better calibrated p -values under the low effect sample size setup. As a concrete example on the scale of experiments, Gasperini et al. (2019) tested about 90,000 perturbation-gene pairs with over 200,000 cells.

The double facets of statistical and computational challenges can also be found in genome-wide association studies (GWAS) even though the source of low effective sample size is different. We postpone the discussion of this example to Section 6. Certainly, the challenges of applying the conditional independence test to modern data analysis can go beyond the low SNR and large scale. In fact, (Shah and Peters, 2020) proved that the conditional independence test (1) is a fundamentally hard problem in the sense that there is no single test that can control Type-I error and simultaneously

have nontrivial power against any alternative without further assumptions.

1.2 Relevant literature

Resampling-based test procedures are particularly suitable for analyzing data with low effective size due to their desirable finite-sample performance, but computation is the main challenge. In fact, in the example of Gasperini et al. (2019), given the multiplicity correction required for such a large number of tests, p -values need to be accurate to about seven decimal places to have a chance at significance (assuming a Bonferroni correction at level $\alpha = 0.05$). Then obtaining accurate p -values for each test requires about a million resamples per p -value, for a total of about 10^{12} resamples across all perturbation-gene pairs tested. In contrast, applying asymptotic tests can result in inflated Type-I error rates in low SNR setup while being much more computationally efficient.

To resolve the dilemma between statistical accuracy and computational efficiency, a number of statistical testing or approximation procedures have been proposed. We divide the discussion of these literatures into two categories: resampling-based and resampling-free procedures. Our discussion will incorporate but go beyond the conditional independence test.

Resampling-based procedures. Resampling-based procedures have been accelerated using adaptive resampling schemes, which adjust the number of resamples drawn based on the data (Besag and Clifford, 1991; Gandy, 2009; Gandy and Hahn, 2014; Gandy and Hahn, 2016; Gandy and Hahn, 2017; Fischer and Ramdas, 2024b; Fischer and Ramdas, 2024a). Such procedures are applicable to arbitrary resampling schemes and test statistics, at the cost of some resampling. On a more applied side, to circumvent the computational barrier of resampling, Ge et al. (2012) and Winkler et al. (2016) proposed heuristic methods to fit parametric curves on the resampled distributions so that the number of resampling can be kept in a relatively small scale used to learn the curves. (Barry et al., 2021) proposed to apply the dCRT (Liu et al., 2022), which is an accelerated variant of the resampling-based *conditional randomization test* (CRT; Candès et al., 2018) in the context of single-cell CRISPR screens, and applied the similar parametric curve fitting approach to the resampling distribution of dCRT.

Resampling-free procedures. A large class of methods in this category are based on finer asymptotic expansion techniques. For example, classical *Edgeworth expansion* has been applied to study the null distribution of test statistics (Hall, 2013; Bentkus, Götze, and Zwet, 1997; Bickel, 1974) to improve upon naive normal approximation. However, the Edgeworth expansion may suffer from accuracy when it comes to the tail area of the statistic distribution. In particular, Chapter 5 in Hall (2013) showed the lack of relative error for Edgeworth expansion estimate, posing a challenge when small p -values are present. *Saddlepoint approximation* (SPA, Daniels (1954) and Lugannani and Rice (1980)) is another classical technique to obtain highly accurate approximations to densities and tail probabilities for quantities that can be expressed as sample averages. In particular, a desirable property of SPA is its relative error guarantee (Butler, 2007; Kolassa, 2006). SPAs have been proposed to approximate resampling dis-

tributions of classical resampling-based procedures, like permutation tests (Robinson, 1982) and the bootstrap (Davison and Hinkley, 1988). Apart from these asymptotic approximations, exact computation of p -value based on the resampled distribution is possible for certain case. In the same paper where the dCRT was introduced (Liu et al., 2022), a resampling-free approximation to this procedure was proposed based on a quantile transformation to a normal distribution. However, these authors acknowledged that this approach is primarily useful for continuously distributed \mathbf{X} , and that it incurs a substantial power loss for discrete \mathbf{X} , the setting that will be focused on in the present work.

With these dense literatures discussed, reconciling the statistical accuracy and computational efficiency remains unresolved and open in the context of conditional independence testing (1).

1.3 Our contributions

In this paper, we propose a new method, the *saddlepoint approximation-based conditional randomization test* (spaCRT). spaCRT is a *resampling-free* conditional independence testing procedure, which reconciles the statistical accuracy of the dCRT with the computational efficiency of asymptotic methods. The key idea is to approximate the distribution of the resampled test statistic in the dCRT using a saddlepoint approximation. Concretely, we make the following contributions:

1. **Establish general approximation accuracy of spaCRT:** In Section 4, we prove the general theoretical results of approximation accuracy of spaCRT in terms of the relative error of p -values and the test is then shown to be asymptotically equivalent to dCRT under mild regularity conditions. Therefore, spaCRT inherits desirable properties of dCRT allowing general regression methods to learn $\mathbb{E}[\mathbf{X}|\mathbf{Z}]$ and $\mathbb{E}[\mathbf{Y}|\mathbf{Z}]$.
2. **Investigate validity of spaCRT with modern regression techniques:** Building on the general results, we show in Section 5 that spaCRT is valid under a wide range of regression techniques, encompassing generalized linear models, high-dimensional regression and nonparametric machine learning regression. The conditions for the accurate approximation of spaCRT towards dCRT is no more stringent than consistency assumptions for the regression estimators.
3. **Explore superior performance in genetics and genomics applications:** In Section 6, we demonstrate the superior finite-sample performance of spaCRT in the simulation setups with sparse outcomes, which are motivated by GWAS and single-cell CRISPR screens. In Section 7, we apply spaCRT to a single-cell CRISPR screens analysis (Gasperini et al., 2019) and show that spaCRT is able to control Type-I error and maintain high power while requiring similar computational resources to the asymptotic tests.
4. **Expand the applicability of SPA in modern statistics:** To the best of our knowledge, spaCRT is the first application of SPA to a conditional independence test. One of the aims of this paper is to draw the attention of the statistical

community to the potential of SPA in modern statistics. Its analytic nature makes SPA particularly well-suited for large-scale data analysis.

Code to reproduce these analyses is available at github.com/Katsevich-Lab/spacrt-manuscript.

1.4 Notation

Define $\text{sgn}(x)$ as the sign of x , i.e. $\text{sgn}(x) = 1$ if $x > 0$, -1 if $x < 0$ and 0 otherwise. For an infinitely differentiable function $f : \mathcal{X} \subset \mathbb{R} \mapsto \mathbb{R}$, define $f^{(r)}$ to be its r -th derivative. Define f' , f'' as the first and second derivative of f respectively. Denote $[n]$ for any $n \in \mathbb{N}_+$ as $\{1, \dots, n\}$. Define $\text{expit}(x) \equiv 1/(1 + \exp(-x))$. Define $\mathbb{E}_{\mathcal{L}_n}[\cdot | \mathcal{F}_n]$, $\mathbb{E}_{\widehat{\mathcal{L}}_n}[\cdot | \mathcal{F}_n]$ as the conditional expectations under law \mathcal{L}_n and its estimate $\widehat{\mathcal{L}}_n$ respectively. Similarly, we define $\text{Var}_{\mathcal{L}_n}[\cdot | \mathcal{F}_n]$ and $\text{Var}_{\widehat{\mathcal{L}}_n}[\cdot | \mathcal{F}_n]$ as the conditional variance counterpart. We use the following standard notations regarding the asymptotic properties of a sequence of random variables X_n :

$$\begin{aligned} X_n = O_{\mathbb{P}}(1) &\quad \text{if for each } \delta > 0 \text{ there is an } M > 0 \text{ s.t. } \limsup_{n \rightarrow \infty} \mathbb{P}[|X_n| > M] < \delta; \\ X_n = \Omega_{\mathbb{P}}(1) &\quad \text{if for each } \delta > 0 \text{ there is an } \eta > 0 \text{ s.t. } \limsup_{n \rightarrow \infty} \mathbb{P}[|X_n| < \eta] < \delta; \\ X_n = o_{\mathbb{P}}(1) &\quad \text{if } \mathbb{P}[|X_n| > \eta] \rightarrow 0 \text{ for all } \eta > 0. \end{aligned}$$

We use $a_n \sim b_n$ if $0 < \liminf_{n \rightarrow \infty} |a_n/b_n| \leq \limsup_{n \rightarrow \infty} |a_n/b_n| < \infty$ and $a_n = o(b_n)$ if $\limsup_{n \rightarrow \infty} |a_n/b_n| \rightarrow 0$.

2 A dilemma between dCRT and GCM

Before introducing the spaCRT, we first discuss the successful application of the dCRT in achieving better Type-I error control compared to its asymptotic counterpart, the *generalized covariance measure* (GCM) test. However, the success in statistical performance builds on the computationally heavy resampling procedure.

2.1 Background: dCRT and GCM test

The dCRT procedure, as proposed by Liu et al. (2022), is designed under the *model-X assumption* that $\mathcal{L}_n(\mathbf{X} | \mathbf{Z})$ is known (Candès et al., 2018). However, this procedure is usually deployed in practice by learning this conditional distribution in-sample. In a prior work, we established the statistical properties of the dCRT with $\mathcal{L}_n(\mathbf{X} | \mathbf{Z})$ estimated in sample (Niu et al., 2024). In this paper, we will refer to the latter procedure as the dCRT, allowing a minor abuse of terminology. Furthermore, we consider the special but still fairly general case when

$$\mathcal{L}_n(\mathbf{X} | \mathbf{Z}) = f(\mathbf{X} | \theta_{n,x}(\mathbf{Z})), \tag{2}$$

where $f(x|\theta)$ is an exponential family with natural parameter θ , natural parameter space \mathbb{R} , and log-partition function A :

$$f(x|\theta) = \exp(\theta x - A(\theta))h(x). \tag{3}$$

This is not a restrictive assumption, since we allow the function $\theta_{n,x}(\mathbf{Z})$ to be arbitrary. Given this setup, consider estimating the functions $\theta_{n,x}(\mathbf{Z})$ and $\mu_{n,y}(\mathbf{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\mathbf{Y} \mid \mathbf{Z}]$ by $\widehat{\theta}_{n,x}(\mathbf{Z})$ and $\widehat{\mu}_{n,y}(\mathbf{Z})$, respectively (we assume throughout that \mathbf{Y} is integrable, so that $\mu_{n,y}$ is well-defined). The learning procedures for these quantities can be arbitrary. The choices include but are not limited to parametric, non-parametric and high-dimensional regression methods. Setting $\widehat{\mu}_{n,x}(\mathbf{Z}) \equiv A'(\widehat{\theta}_{n,x}(\mathbf{Z}))$, we arrive at the test statistic

$$T_n^{\text{dCRT}}(X, Y, Z) = \frac{1}{n} \sum_{i=1}^n (X_{in} - \widehat{\mu}_{n,x}(Z_{in}))(Y_{in} - \widehat{\mu}_{n,y}(Z_{in})). \quad (4)$$

The dCRT is obtained by comparing $T_n^{\text{dCRT}}(X, Y, Z)$ to a null distribution obtained by resampling $X_{in} \mid Z_{in}$ based on the estimated distribution $f(\cdot \mid \widehat{\theta}_{n,x}(Z_{in}))$. This procedure is summarized in Algorithm 1.

Algorithm 1: dCRT procedure with exponential family for $\mathcal{L}_n(\mathbf{X} \mid \mathbf{Z})$

Input: Data (X, Y, Z) , number of randomizations M .

1 Learn $\widehat{\theta}_{n,x}(\cdot)$ and $\widehat{\mu}_{n,x}(\cdot)$ based on (X, Z) ; learn $\widehat{\mu}_{n,y}(\cdot)$ based on (Y, Z) ;

2 Compute $T_n^{\text{dCRT}}(X, Y, Z)$ as in (4);

3 **for** $m = 1, 2, \dots, M$ **do**

4 Sample $\widetilde{X}^{(m)} \mid X, Y, Z \sim \prod_{i=1}^n f(\cdot \mid \widehat{\theta}_{n,x}(Z_{in}))$ and compute

$$T_n^{\text{dCRT}}(\widetilde{X}^{(m)}, X, Y, Z) \equiv \frac{1}{n} \sum_{i=1}^n (\widetilde{X}_{in}^{(m)} - \widehat{\mu}_{n,x}(Z_{in}))(Y_{in} - \widehat{\mu}_{n,y}(Z_{in})); \quad (5)$$

5 **end**

Output: dCRT p -value

$$\frac{1}{M+1} (1 + \sum_{m=1}^M \mathbb{1}\{T_n^{\text{dCRT}}(\widetilde{X}^{(m)}, X, Y, Z) \geq T_n^{\text{dCRT}}(X, Y, Z)\}).$$

dCRT is a resampling-based procedure that can be computationally challenging when the sample size is moderate to large. An alternative test procedure is instead of using the resampling to construct p -value, one can use the normal approximation to the test statistic $T_n^{\text{dCRT}}(X, Y, Z)$ (4) under null and construct a p -value based on the cutoff of the standard normal distribution after properly normalizing the standard deviation estimate. This is the so-called GCM test (Shah and Peters, 2020). Building on the same test statistic (except for extra normalization), GCM also allows flexible modeling choices on estimators $\widehat{\mu}_{n,x}(\cdot)$ and $\widehat{\mu}_{n,y}(\cdot)$. In fact, it has been proved in Shah and Peters (2020) that GCM enjoys the so-called double robustness property: as long as both estimators $\widehat{\mu}_{n,x}(\cdot)$ and $\widehat{\mu}_{n,y}(\cdot)$ are consistent and converge to the true conditional expectations at rate faster than $n^{-1/4}$, the validity of the test can be guaranteed (Shah and Peters, 2020, Theorem 6).

Proved in Niu et al. (2024), the dCRT is asymptotically equivalent to GCM under mild conditions. Thus dCRT inherits the doubly robust statistical property from GCM. One important difference is dCRT relies on resampling to construct the p -value, while GCM relies on the asymptotic normal approximation. This difference can lead to advantage for GCM for analyzing data with large scale because of fast computation.

However, the resampling nature of dCRT (Algorithm 1), though computationally intensive, can lead to improved finite-sample performance. The key insight is that the GCM test, which relies on asymptotic normality, can suffer from slow convergence in the presence of excessive sparsity in the data. In contrast, dCRT leverages resampling, and the resulting distribution of the resampled test statistic can more accurately approximate the sampling distribution of the observed statistic. As an illustrative example, we apply both dCRT and GCM (as well as the proposed spaCRT) to a CRISPR screen analysis in which the perturbation indicators (X) contain many zeros. A comparison of Type-I error and computation time is shown in Figure 1. We observe that dCRT achieves significantly better Type-I error control than the GCM test, although the latter is much faster to compute.

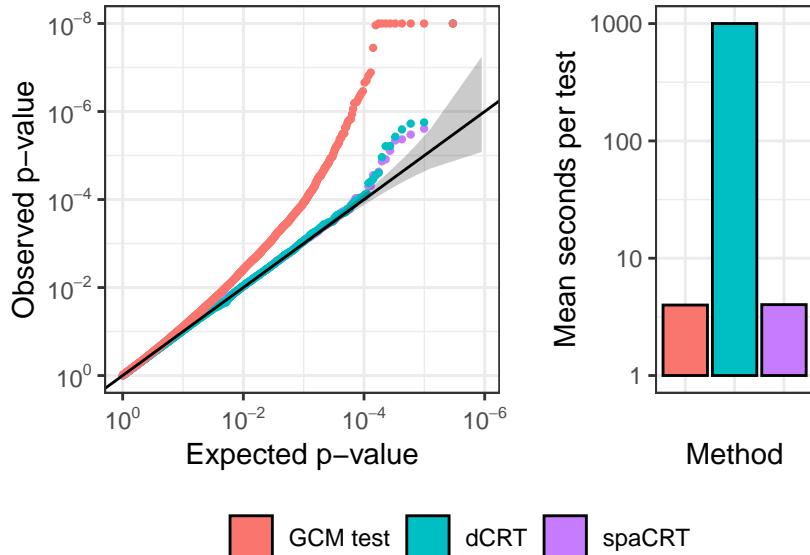


Figure 1: Comparing the Type-I error control and computation times of the GCM test, the dCRT and the proposed spaCRT on the Gasperini et al. (2019) data. We use logistic regression to model the perturbation $\mathbf{X}|\mathbf{Z}$ and negative binomial regression to model the outcome $\mathbf{Y}|\mathbf{Z}$. Left: QQ-plot of the p -values under the null hypothesis, obtained from testing 51 *non-targeting* perturbations against 3,000 genes. The p -values are truncated from below at 10^{-8} for visualization purposes. Right: Mean computation times per perturbation-gene pair, in seconds.

The improvement of dCRT upon GCM can be justified theoretically with excessive sparsity in the data. We will use an illustrative result to show how GCM can suffer from the slow convergence rate of Type-I error while dCRT preserves the Type-I error control in the next section.

2.2 Illustrative results with oracle knowledge of $\mu_{n,x}$ and $\mu_{n,y}$

To develop the intuition, we work with a simple but illustrative setup where we assume that the conditional expectations $\mu_{n,x}(\cdot)$ and $\mu_{n,y}(\cdot)$ are known and we will set $\hat{\mu}_{n,x}(\cdot) = \mu_{n,x}(\cdot)$ and $\hat{\mu}_{n,y}(\cdot) = \mu_{n,y}(\cdot)$ in the test statistics. We will focus on a Bernoulli model

for $\mathbf{X}|\mathbf{Z}$:

$$\mathbf{X}|\mathbf{Z} \sim \text{Ber}(\mu_{n,x}(\mathbf{Z})). \quad (6)$$

We define the *oracle* GCM (oGCM) test by considering the test statistic

$$\phi_{n,\alpha}^{\text{oGCM}} \equiv \mathbb{1}(T_n^{\text{oGCM}}(X, Y, Z) > z_{1-\alpha}) \quad \text{where} \quad T_n^{\text{oGCM}}(X, Y, Z) \equiv \frac{1}{nS_n} \sum_{i=1}^n R_{in}^o, \quad (7)$$

$R_{in}^o \equiv (X_{in} - \mu_{n,x}(Z_{in}))(Y_{in} - \mu_{n,y}(Z_{in}))$ and $S_n^2 = \mathbb{E}[(R_{in}^o)^2]$. For dCRT, we consider the modified dCRT with theoretical quantile, which we call oracle dCRT (odCRT):

$$\phi_{n,\alpha}^{\text{odCRT}} \equiv \mathbb{1}\left(T_n^{\text{dCRT}} \geq \mathbb{Q}_{1-\alpha}(\tilde{T}_n^{\text{dCRT}}|X, Y, Z)\right)$$

where T_n^{dCRT} , $\tilde{T}_n^{\text{dCRT}}$ are defined in (4) and (5) respectively, but with oracle $\mu_{n,x}(\cdot)$, $\mu_{n,y}(\cdot)$ and resampling distribution $\tilde{X}^{(m)} \sim \prod_{i=1}^n \text{Ber}(\mu_{n,x}(Z_{in}))$. The intuition for the Type-I error deviating from the specified significance level for GCM (and oGCM) is the CLT, when excessive sparsity exists, can happen in an arbitrarily slow rate depending how sparse the data is. To formalize such intuition, we consider the following assumptions.

Assumption 1 (Sparsity level in \mathbf{X}). *Suppose v_n is a sequence of positive constants and $cv_n \leq \inf_z |\mu_{n,x}(z)| \leq \sup_z |\mu_{n,x}(z)| \leq Cv_n$ for some universal constants $C > 0$ and $c > 0$.*

Assumption 2 (Conditional moments of \mathbf{Y}). *Suppose the following conditions hold: $\sup_z \mathbb{E}[\mathbf{Y}^4|\mathbf{Z} = z] < \infty$, $\inf_z \mathbb{E}[(\mathbf{Y} - \mathbb{E}[\mathbf{Y}|\mathbf{Z}])^3|\mathbf{Z} = z] > 0$ and $\inf_z \text{Var}[\mathbf{Y}|\mathbf{Z} = z] > 0$.*

Assumption 3 (Cramér's condition). *Suppose $S_{in} \equiv R_{in}^o / \sqrt{\mathbb{E}[(S_n)^2]}$ satisfies the Cramér's condition: $\limsup_{n \rightarrow \infty} \sup_{t \in \mathbb{R}} |\mathbb{E}[\exp(itS_{in})]| < 1$.*

Parameter v_n in Assumption 1 characterizes the sparsity level of data, which will play an important role in Theorem 1 to unveil the failure of oGCM to control Type-I error under small sample size. Assumption 2 states the bounded moment condition for \mathbf{Y} given \mathbf{Z} as well as the non-degeneracy of the conditional variance and conditional third central moment. This is mainly required to prove the rate of convergence on Type-I error for oGCM test. Such assumption can be satisfied by examples including Poisson or negative binomial case with uniformly lower and upper bounded conditional mean (and fixed dispersion parameter for negative binomial case) in $\mathbf{Y}|\mathbf{Z}$. This corresponds to the setup in Figure 1. Assumption 3 is used to guarantee the validity of *Edgeworth expansion* on S_{in} . The assumption may seem to be contradictory with model setup (6) at the first glance because of potential sparsity in X . However, this is not the case. The key reason hinges on the convolution nature of random variable S_{in} and as long as $\mu_{n,x}(Z_{in}) \cdot \mu_{n,y}(Z_{in})$ are continuous random variables, the convolution of the product variable with discrete random variables can still satisfy the Cramér's condition. Now we state our illustrative results.

Theorem 1. *Consider $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$. Suppose Assumption 1-3 hold. Then we have*

1. **Finite-sample validity of odCRT:** $\mathbb{E}[\phi_{n,\alpha}^{\text{odCRT}}] = \alpha$;

2. **Convergence of Type-I error of oGCM:** If $1/v_n = o(n)$, then there exists a sequence $r_n > 0$ such that $r_n \sim 1/(nv_n)^{1/2}$ and

$$|\mathbb{E}[\phi_{n,\alpha}^{\text{oGCM}}] - \alpha - r_n| = o(r_n). \quad (8)$$

The argument for finite-sample validity is by the exchangeability of the resampled data and the original data under null hypothesis $H_0 : \mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$. Now we discuss the implication of the results on oGCM.

Remark 1 (Implication for testing with oGCM). Theorem 1 unveils that sparsity in data can slow down the Type-I error converging to the specified significance level α , which is a result of slow convergence rate of CDF of oGCM test statistic converging to the CDF of standard normal distribution. When v_n is of order n^{-s} for $s > 0$, the convergence rate of Type-I error of oGCM is $n^{(1-s)/2}$. The closer s is to 1, the slower the convergence rate is.

Theorem 1 considers the model-X assumption for odCRT and oracle knowledge of $\mu_{n,x}, \mu_{n,y}$. Thus it only serves as a illustration and the results are not directly applicable to the general case where $\mu_{n,x}(\cdot)$ and $\mu_{n,y}(\cdot)$ are unknown. However, the theorem provides a high-level insight on the finite-sample performance of dCRT and GCM tests.

3 spaCRT: A resampling-free approximation to dCRT

While the improvement of dCRT over GCM is quite evident both from empirical example and rigorous theoretical result as illustrated in Section 2, the computational burden of dCRT is still a critical concern. The resampling procedure in Algorithm 1 is slow and can be infeasible for large datasets. In this section, we propose a new test, the *saddlepoint approximation-based conditional randomization test* (spaCRT), which is completely resampling-free and has similar finite-sample performance as dCRT. The computation gain with spaCRT is significant and obtained without losing the statistical accuracy (Figure 1).

3.1 The spaCRT

If we consider the limit of the dCRT p -value as the number of resamples M grows indefinitely, we obtain

$$p_{\text{dCRT}} \equiv \mathbb{P} \left[T_n^{\text{dCRT}}(\tilde{X}, X, Y, Z) \geq T_n^{\text{dCRT}}(X, Y, Z) \mid X, Y, Z \right].$$

We approximate this conditional tail probability via the SPA. Note that the resampled test statistic defined in (5) is the mean of conditionally independent random variables:

$$T_n^{\text{dCRT}}(\tilde{X}^{(m)}, X, Y, Z) \equiv \frac{1}{n} \sum_{i=1}^n W_{in}, \quad W_{in} \equiv a_{in}(\tilde{X}_{in} - \hat{\mu}_{n,x}(Z_{in})), \quad a_{in} \equiv Y_{in} - \hat{\mu}_{n,y}(Z_{in}).$$

Indeed, W_{in} are independent, but not identically distributed, conditionally on the σ -algebra $\mathcal{F}_n \equiv \sigma(X, Y, Z)$. In a parallel work (Niu, Ray Choudhury, and Katsevich,

2024), we have established an SPA result for means of conditionally independent random variables under relatively mild conditions. This result is restated here as Lemma 4 in Appendix B. This result is expressed in terms of the average conditional cumulant-generating function

$$K_n(s \mid \mathcal{F}_n) \equiv \frac{1}{n} \sum_{i=1}^n K_{in}(s \mid \mathcal{F}_n) \equiv \frac{1}{n} \sum_{i=1}^n \log \mathbb{E}[\exp(sW_{in}) \mid \mathcal{F}_n], \quad (9)$$

which in our case can be expressed as

$$K_n(s \mid \mathcal{F}_n) = \frac{1}{n} \sum_{i=1}^n \left\{ A(\widehat{\theta}_{n,x}(Z_{in}) + a_{in}s) - A(\widehat{\theta}_{n,x}(Z_{in})) - a_{in}sA'(\widehat{\theta}_{n,x}(Z_{in})) \right\}. \quad (10)$$

The first two derivatives of this quantity are

$$K'_n(s \mid \mathcal{F}_n) = \frac{1}{n} \sum_{i=1}^n a_{in} \left(A'(\widehat{\theta}_{n,x}(Z_{in}) + a_{in}s) - A'(\widehat{\theta}_{n,x}(Z_{in})) \right), \quad (11)$$

$$K''_n(s \mid \mathcal{F}_n) = \frac{1}{n} \sum_{i=1}^n a_{in}^2 A''(\widehat{\theta}_{n,x}(Z_{in}) + a_{in}s). \quad (12)$$

We can now present the spaCRT procedure (Algorithm 2).

Algorithm 2: spaCRT procedure

Input: Data (X, Y, Z) .

- 1 Learn $\widehat{\theta}_{n,x}(\cdot)$ and $\widehat{\mu}_{n,x}(\cdot)$ based on (X, Z) , $\widehat{\mu}_{n,y}(\cdot)$ based on (Y, Z) ;
- 2 Compute $T_n^{\text{dCRT}}(X, Y, Z)$ as in (4);
- 3 Find \hat{s}_n that solves the saddlepoint equation

$$K'_n(s \mid \mathcal{F}_n) = T_n^{\text{dCRT}}(X, Y, Z); \quad (13)$$

- 4 Compute $\lambda_n = \sqrt{n}\hat{s}_n\sqrt{K''_n(\hat{s}_n \mid \mathcal{F}_n)}$ and

$$r_n = \begin{cases} \text{sgn}(\hat{s}_n)\sqrt{2n(\hat{s}_n T_n^{\text{dCRT}} - K_n(\hat{s}_n \mid \mathcal{F}_n))}, & \text{if } \hat{s}_n T_n^{\text{dCRT}} - K_n(\hat{s}_n \mid \mathcal{F}_n) \geq 0; \\ \text{sgn}(\hat{s}_n) & \text{otherwise.} \end{cases}$$

Output: spaCRT p -value

$$p_{\text{spaCRT}} \equiv 1 - \Phi(r_n) + \phi(r_n) \left\{ \frac{1}{\lambda_n} - \frac{1}{r_n} \right\}. \quad (14)$$

The spaCRT procedure is attractive because it is completely resampling-free. It requires the following one-time computations: fitting the estimates $\widehat{\theta}_{n,x}$ and $\widehat{\mu}_{n,y}$, calculating the test statistic T_n^{dCRT} , and finding the solution to the saddlepoint equation (13). The latter is a one-dimensional root-finding problem and can be solved efficiently using standard numerical optimization algorithms.

To make the spaCRT procedure more concrete, we provide an example in the case that \mathbf{X} is binary, a setting that matches our motivating application.

Example 1 (Bernoulli sampling). Suppose $\mathbf{X} \mid \mathbf{Z} \sim \text{Ber}(\mu_{n,x}(\mathbf{Z}))$, and $\theta_{n,x}(\mathbf{Z}) = \text{logit}(\mu_{n,x}(\mathbf{Z}))$. Then, we have $A(\theta) = \log(1 + \exp(\theta))$. After some manipulation, the saddlepoint equation reduces to

$$\frac{1}{n} \sum_{i=1}^n (Y_{in} - \hat{\mu}_{n,y}(Z_{in}))(X_{in} - \text{expit}(\hat{\theta}_{n,x}(Z_{in}) + s(Y_{in} - \hat{\mu}_{n,y}(Z_{in})))) = 0.$$

Defining $\tilde{\mu}_{n,x}(Z_{in}) \equiv \text{expit}(\hat{\theta}_{n,x}(Z_{in}) + \hat{s}_n(Y_{in} - \hat{\mu}_{n,y}(Z_{in})))$ for convenience, λ_n and r_n can be computed as

$$\lambda_n = \hat{s}_n \sqrt{\sum_{i=1}^n (Y_{in} - \hat{\mu}_{n,y}(Z_{in}))^2 \tilde{\mu}_{n,x}(Z_{in})(1 - \tilde{\mu}_{n,x}(Z_{in}))}$$

and

$$r_n = \text{sgn}(\hat{s}_n) \sqrt{2 \sum_{i=1}^n \left(X_{in} \log \frac{\tilde{\mu}_{n,x}(Z_{in})}{\hat{\mu}_{n,x}(Z_{in})} + (1 - X_{in}) \log \frac{1 - \tilde{\mu}_{n,x}(Z_{in})}{1 - \hat{\mu}_{n,x}(Z_{in})} \right)},$$

or simply $\text{sgn}(\hat{s}_n)$ if the quantity under the square root is negative. Putting these pieces together, the spaCRT p -value can be computed as in equation (14).

4 General theory of the spaCRT

spaCRT does not require any resampling and thus has a significant advantage over dCRT in terms of computation. This advantage does not come with a sacrifice of statistical accuracy. In this section, we establish the theoretical properties of the spaCRT. We first state general results concerning the approximation accuracy of the spaCRT p -value and then prove the asymptotic equivalence between the spaCRT and dCRT.

Theorem 2 (Approximation accuracy). *Suppose there exists $S > 0$ such that one of the following conditions holds:*

$$\sup_i |\hat{\theta}_{n,x}(Z_{in})|, \sup_i |\hat{\mu}_{n,y}(Z_{in})| = O_{\mathbb{P}}(1), \mathbb{P}[Y_{in} \in [-S, S]] = 1 \text{ for any } i, n; \quad (\text{CSE})$$

$$\frac{1}{n} \sum_{i=1}^n (Y_{in} - \hat{\mu}_{n,y}(Z_{in}))^4 = O_{\mathbb{P}}(1), \mathbb{P}\left[\tilde{X}_{in} \in [-S, S]\right] = 1 \text{ for any } i, n. \quad (\text{CCS})$$

Suppose the following conditions hold:

$$|\hat{\theta}_{n,x}(Z_{in})| < \infty, |\hat{\mu}_{n,y}(Z_{in})| < \infty \text{ for any } i, n \text{ almost surely}; \quad (15)$$

$$\frac{1}{n} \sum_{i=1}^n (Y_{in} - \hat{\mu}_{n,y}(Z_{in}))^2 A''(\hat{\theta}_{n,x}(Z_{in})) = \Omega_{\mathbb{P}}(1); \quad (16)$$

$$T_n^{\text{dCRT}}(X, Y, Z) \xrightarrow{\mathbb{P}} 0. \quad (17)$$

Then, the saddlepoint equation (13) has a unique and finite solution $\hat{s}_n \in [-1/16, 1/16]$ with probability approaching 1 as $n \rightarrow \infty$. Furthermore, the spaCRT p -value p_{spaCRT} approximates the dCRT p -value p_{dCRT} with vanishing relative error:

$$p_{\text{dCRT}} = p_{\text{spaCRT}} \cdot (1 + o_{\mathbb{P}}(1)) \quad (18)$$

and spaCRT p -value is positive with probability approaching 1 as $n \rightarrow \infty$:

$$\mathbb{P}[p_{\text{spaCRT}} > 0] \rightarrow 1 \text{ as } n \rightarrow \infty. \quad (19)$$

To better understand the assumptions in Theorem 2, we provide the following remarks.

Remark 2 (Comments on the assumptions). Assumptions (CSE)-(CCS) are conditions that can be satisfied if the random variables involved have light enough tails and estimators $\widehat{\theta}_{n,x}(Z_{in}), \widehat{\mu}_{n,y}(Z_{in})$ are regular enough. Assumptions (15) and (16) are mild; their purpose is to rule out degenerate cases. Finally, the role of the assumption (17) is to guarantee the existence of the solution to the saddlepoint equation. This assumption allows the test statistic $T_n^{\text{dCRT}}(X, Y, Z)$ to converge to zero in probability, **at any rate**. In particular, we consider the following two most important cases among others:

1. **Under the null hypothesis:** Shah and Peters (2020) proved that under general conditions on $\widehat{\mu}_{n,x}, \widehat{\mu}_{n,y}$, $n^{1/2}T_n^{\text{dCRT}}(X, Y, Z)$ converges weakly to a normal distribution under the null hypothesis. Thus the condition (17) is satisfied under the null hypothesis with rate $n^{-1/2}$.
2. **Under contiguous local alternatives:** The proof of Theorem 3 in Niu et al. (2024) shows that under generalized partially linear models, the test statistic $n^{1/2}T_n^{\text{dCRT}}(X, Y, Z)$ converges to a normal distribution with nonzero mean and positive finite variance under local alternatives that are contiguous to the null distribution. Thus the condition is satisfied in this case with rate $n^{-1/2}$.

Remark 3 (Relative error guarantee). The relative error guarantee in conclusion (18) is a strong result. It means not only the difference of p -values is close to 0 with probability approaching 1, but also the ratio of p -values is close to 1 with probability approaching 1. This is a particularly desirable property for approximating small p -values.

It is not hard to believe the equivalence of dCRT and spaCRT can be easily drawn given the approximate nature of spaCRT towards dCRT. We can formalize such intuition into rigorous results. To proceed with the statement of the results, we define the level- α tests associated with the dCRT and spaCRT p -values:

$$\phi_{n,\alpha}^{\text{dCRT}} \equiv \mathbb{1}(p_{\text{dCRT}} \leq \alpha) \quad \text{and} \quad \phi_{n,\alpha}^{\text{spaCRT}} \equiv \mathbb{1}(p_{\text{spaCRT}} \leq \alpha).$$

The following theorem states that these two tests are asymptotically equivalent.

Theorem 3. Suppose the assumptions of Theorem 2 hold. Fix $\alpha \in (0, 1)$. If the normalized test statistic $n^{1/2}T_n^{\text{dCRT}}(X, Y, Z)/\widehat{S}_n^{\text{dCRT}}$, where $\widehat{S}_n^{\text{dCRT}}$ is defined in equation

(44) (in Appendix), does not accumulate around the $1 - \alpha$ quantile of standard normal distribution $z_{1-\alpha}$, i.e.,

$$\lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \mathbb{P}_{\mathcal{L}_n} \left[\left| \frac{n^{1/2} T_n^{\text{dCRT}}(X, Y, Z)}{\widehat{S}_n^{\text{dCRT}}} - z_{1-\alpha} \right| \leq \delta \right] = 0, \quad (20)$$

then the dCRT and spaCRT tests are asymptotically equivalent:

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\mathcal{L}_n} [\phi_{n,\alpha}^{\text{spaCRT}} = \phi_{n,\alpha}^{\text{dCRT}}] = 1 \quad \text{and} \quad \lim_{n \rightarrow \infty} \mathbb{E}_{\mathcal{L}_n} [\phi_{n,\alpha}^{\text{spaCRT}}] - \mathbb{E}_{\mathcal{L}_n} [\phi_{n,\alpha}^{\text{dCRT}}] = 0.$$

Furthermore, the asymptotic Type-I error control of the spaCRT follows from that of the dCRT even without the regularity condition (20):

Corollary 1 (Asymptotic validity of spaCRT). *Suppose the assumptions of Theorem 2 hold. Fix $\alpha \in (0, 1)$. If $\lim_{n \rightarrow \infty} \mathbb{P}_{H_0}[p_{\text{dCRT}} \leq \alpha] \leq \alpha$, then we have*

$$\lim_{n \rightarrow \infty} \mathbb{P}_{H_0}[p_{\text{spaCRT}} \leq \alpha] \leq \alpha.$$

Remark 4 (Asymptotic validity of spaCRT). Corollary 1 states the asymptotic validity of spaCRT given the asymptotic validity of dCRT. dCRT is proposed originally assuming the exact knowledge of $\mathcal{L}_n(\mathbf{X} | \mathbf{Z})$, or the so-called Model-X assumption. Under the model-X assumptions, the p -value produced by dCRT procedure is exact, i.e., $\mathbb{P}_{H_0}[p_{\text{dCRT}} \leq \alpha] \leq \alpha$. Therefore, p -value obtained from spaCRT procedure is asymptotically valid under the assumptions of Theorem 2. For asymptotic validity of dCRT with general in-sample fit $\widehat{\mathcal{L}}_n(\mathbf{X} | \mathbf{Z})$, we refer reader to the detailed discussion in Niu et al. (2024).

5 Case studies with modern regression techniques

We have established the approximation accuracy of the spaCRT towards dCRT under general conditions. We will dedicate the following three sections to special cases where spaCRT can be provably valid under null hypothesis and thus quite useful for the conditional independence testing with complex data structure. Before proceeding to the particular case studies, we need the following assumptions to state the formal results.

Assumption 4. $0 < \inf_n \mathbb{E}[(X_{in} - \mathbb{E}[X_{in} | Z_{in}])^2(Y_{in} - \mathbb{E}[Y_{in} | Z_{in}])^2]$.

Assumption 5. *Support of $\mathbf{Z} \in \mathbb{R}^d$ is compact, i.e., there exists C_Z such that $\|\mathbf{Z}\|_\infty \leq C_Z < \infty$.*

Throughout this section, we will consider the following logistic model for $\mathbf{X} | \mathbf{Z}$:

$$\mathbf{X} | \mathbf{Z} \sim \text{Ber}(\text{expit}(\mathbf{Z}^\top \boldsymbol{\theta}_n)).$$

The choice of logistic model is mainly for the ease of the theoretical analysis. In fact, even for such binary valued \mathbf{X} , the spaCRT can be easily integrated with fitting

methods beyond just logistic regression, for example hidden Markov models as illustrated in the simulations in Section 6.2 or nonparametric random forest classification in Appendix M.

In the following sections, we investigate the performance of spaCRT when $\mathbf{Y}|\mathbf{Z}$ is estimated using modern regression techniques encompassing low-dimensional generalized linear regression, high-dimensional regression and nonparametric machine learning regression. In particular, we will consider the maximum likelihood estimate (MLE) for the low-dimensional case, lasso regression for the high-dimensional regression and kernel ridge regression for the nonparametric regression.

5.1 Low-dimensional generalized linear regression

Throughout this section, we will assume we are under the classical low-dimensional setup so that we can simplify the subscript $(Z_{in}, X_{in}, Y_{in}) = (Z_i, X_i, Y_i)$ and $\theta_n = \theta$. To echo the simulation result in Figure 1, we consider the conditional distribution $\mathbf{Y} | \mathbf{Z}$ to be:

$$\mathbf{Y} | \mathbf{Z} \sim \mathcal{P}_y(A'_y(\mathbf{Z}^\top \beta))$$

where \mathcal{P}_y is a distribution in the natural exponential family (3) and A_y is the log-parition functions for $\mathbf{Y} | \mathbf{Z}$. Now we state our result.

Theorem 4. *Consider $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$. Suppose assumptions 4-5 hold. Then if the MLEs $\hat{\beta}, \hat{\theta}$ satisfy the following assumptions hold:*

$$\|\hat{\beta} - \beta\|_1 = O_{\mathbb{P}}(1/\sqrt{n}) \quad \text{and} \quad \|\hat{\theta} - \theta\|_1 = O_{\mathbb{P}}(1/\sqrt{n}) \quad (21)$$

Then the conclusion in Theorem 2 holds and we have $\phi_{n,\alpha}^{\text{spaCRT}}$ is asymptotically valid, i.e. $\lim_{n \rightarrow \infty} \mathbb{E}[\phi_{n,\alpha}^{\text{spaCRT}}] = \alpha$.

We want to note that the \sqrt{n} rate condition in (21) is classical for MLEs.

5.2 High-dimensional regression

In this section, we establish the validity of the spaCRT under the following generalized linear models for $\mathbf{Y} | \mathbf{Z}$:

$$\mathbf{Y} | \mathbf{Z} \sim \mathcal{P}_y(A'_y(\mathbf{Z}^\top \beta_n)),$$

\mathcal{P}_y is a distribution in the natural exponential family (3) and A_y is the log-parition functions for $\mathbf{Y} | \mathbf{Z}$. Note this setup is similar to that in Section 5.1 while here we allow dimension of \mathbf{Z} (as well as β_n) to grow with sample size n . We will demonstrate how spaCRT can be used in this setup and thus we need to compute the estimators $\hat{\mu}_{n,y}(\mathbf{Z})$ for $\mathbb{E}[\mathbf{Y} | \mathbf{Z}]$ and $\hat{\mu}_{n,x}(\mathbf{Z})$ for $\mathbb{P}[\mathbf{X} = 1 | \mathbf{Z}] = \mathbb{E}[\mathbf{X} | \mathbf{Z}]$ as required in Algorithm 2. In particular, we consider the estimators $\hat{\beta}_n, \hat{\theta}_n$ for β_n, θ_n are obtained from the following lasso estimators (Tibshirani, 1996):

$$\hat{\beta}_n = \arg \min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^n \{A_y(Z_{in}^\top \beta) - Y_{in} \cdot (Z_{in}^\top \beta)\} + \lambda_n \|\beta\|_1 \right\} \quad (22)$$

and

$$\hat{\theta}_n = \arg \min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^n \{ \log(1 + \exp(Z_{in}^\top \theta)) - X_i \cdot (Z_{in}^\top \theta) \} + \nu_n \|\theta\|_1 \right\}. \quad (23)$$

Recovering the true parameters β_n, θ_n is a challenging (sometimes even unidentifiable) task in the high-dimensional setting unless certain structure assumptions on these parameters are imposed. We show that the spaCRT is valid under the sparse signal setup. We start by stating the following assumption on the covariate distribution \mathbf{Z} .

Assumption 6 (Design assumption). *Suppose the distribution of \mathbf{Z} satisfies the following conditions.*

$$\text{minimum eigenvalue of } \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top] \text{ is bounded away from 0}; \quad (24)$$

$$\mathbb{E}[\langle \boldsymbol{\eta}, \mathbf{Z} \rangle^4] \leq \kappa < \infty, \forall \boldsymbol{\eta} \in \mathbb{R}^d, \|\boldsymbol{\eta}\|_2 = 1. \quad (25)$$

Theorem 5. *Consider $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$. Suppose Assumptions 4-6 hold and there exists $\delta \in (0, 1)$ such that:*

$$\max \{1, s_{\theta_n}, s_{\beta_n}\} \sqrt{\log(d)/n} \sim n^{-\delta} \text{ where } (s_{\beta_n}, s_{\theta_n}) \equiv (\|\beta_n\|_0, \|\theta_n\|_0); \quad (26)$$

$$\sup_n \|\theta_n\|_1 < \infty, \sup_n \|\beta_n\|_1 < \infty. \quad (27)$$

Then if we choose $\lambda_n = C_\lambda \sqrt{\log(d)/n}$ and $\nu_n = C_\nu \sqrt{\log(d)/n}$ for some universal constants C_λ, C_ν , then conclusion in Theorem 2 hold. If additionally, we have

$$s_{\theta_n} \cdot s_{\beta_n} \cdot \frac{\log(d)}{n^{1/2}} = o(1) \quad (28)$$

then $\phi_{n,\alpha}^{\text{spaCRT}}$ is asymptotically valid, i.e. $\lim_{n \rightarrow \infty} \mathbb{E}[\phi_{n,\alpha}^{\text{spaCRT}}] = \alpha$ for any $\alpha \in (0, 1)$.

The conditions required in Theorem 5 are sufficient to prove the strong consistency of the estimators $\hat{\beta}_n, \hat{\theta}_n$ towards the population quantities. This will serve as basic building blocks for providing sufficient conditions for Theorem 2 and also the validity of spaCRT when working under high-dimensional regression setup. In particular, Assumption 6 imposes the structures on the covariate distribution which are commonly required in the high-dimensional regression. In fact, conditions (24)-(25) are also required in Example 9.17 and Theorem 9.36 in Wainwright (2019) when proving the *restricted strong convexity*, a fundamental property involved when proving consistency results of high-dimensional regression coefficients (see for example Corollary 9.26 in Wainwright (2019)). Condition (26) regulates the rate growth between the sparsity parameters, the dimensionality of the covariate and the sample size and is important when proving the almost sure convergence (strong consistency) of the lasso estimators. The boundedness condition (27) is a relatively mild condition required by showing the almost sure convergence of $\hat{\mu}_{n,y}(\cdot), \hat{\theta}_{n,x}(\cdot)$ when verifying condition (15) recalling the definition of $\theta_{n,x}$ and its estimate $\hat{\theta}_{n,x}$ in Section 2.1.

5.3 Nonparametric machine learning regression

In this section, we study the validity of spaCRT when the conditional distribution $\mathbf{Y}|\mathbf{Z}$ is modeled using *kernel ridge regression* (KRR), a representative of nonparametric machine learning methods. Throughout this section, we will assume we are under the classical low-dimensional setup so that we can simplify the subscript $(Z_{in}, X_{in}, Y_{in}) = (Z_i, X_i, Y_i)$ and $\theta_n = \theta$.

Suppose the conditional expectations $\mu_{n,y} \in \mathcal{H}$ for some RKHS $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ with reproducing kernel $k \in \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. Let $K \in \mathbb{R}^{n \times n}$ have ij th entry $K_{ij} = k(Z_i, Z_j)/n$ and denote the eigenvalues of K by $\widehat{\kappa}_1 \geq \widehat{\kappa}_2 \geq \dots \geq \widehat{\kappa}_n \geq 0$. We will assume that kernel function k admits an eigen-expansion of the form

$$k(z, z') = \sum_{j=1}^{\infty} \kappa_j e_j(z) e_j(z') \quad (29)$$

with orthonormal eigenfunctions $\{e_j\}_{j=1}^{\infty}$, so $\mathbb{E}[e_j e_k] = \mathbb{1}(k=j)$, and summable eigenvalues $\kappa_1 \geq \kappa_2 \geq \dots \geq 0$. Such expansion can be guaranteed by Mercer's theorem if mild conditions are satisfied. For a sequence of regularization parameter λ_n , we consider the following estimator:

$$\widehat{\mu}_y \equiv \arg \min_{\mu_y \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - \mu_y(Z_i))^2 + \lambda_n \|\mu_y\|_{\mathcal{H}}^2 \right\}. \quad (30)$$

We consider selecting λ_n in the following data-dependent way:

$$\lambda_n = \arg \min_{\lambda > 0} \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\widehat{\kappa}_i^2}{(\widehat{\kappa}_i + \lambda)^2} + \lambda \right\} \quad (31)$$

We want to emphasize that the way we select the tuning parameter is mainly for the ease of theoretical analysis and similar data-dependent hyperparameter selection has been adopted in previous work Niu et al. (2024) and Shah and Peters (2020). As for the estimator $\widehat{\theta}$ for θ , we consider using the maximum likelihood estimator $\widehat{\theta}$. With the estimators $\widehat{\mu}_y(Z_i)$ and $\widehat{\theta}_{n,x}(Z_i) = Z_i^\top \widehat{\theta}$, the spaCRT can be applied with Algorithm 2. Now we state our main results on the validity guarantee of spaCRT.

Theorem 6. *Consider $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$. Suppose assumptions 4-5 hold. Then if the following conditions hold:*

$$\text{support of } \mathbf{Y} \text{ is compact, i.e., there exists } S, \mathbb{P}[\mathbf{Y} \in [-S, S]] = 1; \quad (32)$$

$$\|\widehat{\theta} - \theta\|_1 = O_{\mathbb{P}}(1/\sqrt{n}); \quad (33)$$

$$\|\widehat{\mu}_y\|_{\infty} = O_{\mathbb{P}}(1); \quad (34)$$

$$\sum_{j=1}^{\infty} \kappa_j < \infty, \quad (35)$$

then the conclusion in Theorem 2 hold and $\lim_{n \rightarrow \infty} \mathbb{E}[\phi_{n,\alpha}^{\text{spaCRT}}] = \alpha$.

Conditions in Theorem 6 are mild conditions. Condition (34) can be easily verified when linear kernel, i.e. linear ridge regression, is considered. For general choice of kernel, condition (34) can hold under extra conditions on the kernel function k .

6 Application to data with excessive sparsity

As discussed in Section 1, the spaCRT method can be very useful when it comes to analysis with very sparse data, i.e. excessive zeros in outcome data Y and sometimes in treatments X . Such phenomena are ubiquitous in modern dataset, especially those generated in biological studies due to either technical machine noise or the low signal-to-noise (SNR) nature of biological processes. In order to demonstrate the advantage of using the spaCRT method in the data with sparsity, we consider two simulation examples, single-cell CRISPR screens analysis and GWAS with rare diseases, discussed in Section 1. We will focus on showcasing these two examples in the remaining parts of this section. In addition to these two examples, we will also provide another example in Appendix M inspired from unblanced classification in machine learning.

In the following sections, we will first discuss the source of sparsity in the context where data is collected and then we discuss the methods compared, the simulation results and the interpretation of these results.

6.1 Single-cell CRISPR screens analysis

Source of sparsity in single-cell CRISPR screens: Recall in such experiments, each cell receives several perturbations targeting different genome elements. Due to such pooling of a large number of perturbations in a single experiment, most perturbations are present in only a small fraction of cells. Furthermore, gene expression data are measured as RNA molecule counts, and when measured at single-cell resolution, the relatively small number of total RNA molecules measured per cell and the large number of genes result in many genes having zero expression in most cells (Svensson, 2020).

Simulation setup: To mimic the application of single-cell CRISPR screens, we model $\mathbf{X} | \mathbf{Z}$ as a logistic regression model and $\mathbf{Y} | \mathbf{Z}$ as a negative binomial regression model (Barry et al., 2024; Barry et al., 2021; Gasperini et al., 2019). The latter modeling choice is quite common not just in single-cell CRISPR screen analysis but in single-cell RNA-seq analysis more broadly (Huang et al., 2018; Townes et al., 2019; Svensson, 2020). We consider the following data-generating model:

$$\mathbf{Z} \sim N(0, 1); \quad \mathbf{X} | \mathbf{Z} \sim \text{Ber}(\text{expit}(\gamma_0 + \mathbf{Z})); \quad \mathbf{Y} | \mathbf{X}, \mathbf{Z} \sim \text{NB}(\text{exp}(\beta_0 + \rho \mathbf{X} + \mathbf{Z}), r), \quad (36)$$

where $r > 0$ is the *size parameter* controlling the overdispersion of the negative binomial distribution. Here, \mathbf{X} , \mathbf{Y} , and \mathbf{Z} represent the indicator of perturbation presence, gene expression, and a single covariate with a confounding effect, respectively. The parameters γ_0 and β_0 control the proportion of cells with perturbations and the mean expression of the gene, respectively, and therefore control the sparsity level of X and Y . The smaller γ_0 and β_0 , the sparser X and Y and the range of these parameters are chosen to roughly match the sparsity level in the real data analyzed in the next section. In this setup, we consider testing for association between a single CRISPR perturbation and a single gene. The parameter ρ controls the strength of the signal, i.e., the dependence of \mathbf{Y} on \mathbf{X} conditional on \mathbf{Z} . Therefore, $\rho = 0$ and $\rho \neq 0$ corresponds to the null and alternative hypotheses, respectively.

Methodologies compared: We compare the following four tests.

- The **spaCRT** (Algorithm 2), where $\mathbf{X} \mid \mathbf{Z}$ is fit based on a logistic regression model and $\mathbf{Y} \mid \mathbf{Z}$ is fit based on a negative binomial regression model.
- The **dCRT** (Algorithm 1), with the same fitting procedures as the spaCRT and $M = 10,000$.
- The **GCM test** (Shah and Peters, 2020), which is based on the asymptotically normal test statistic, a normalized version of $T_n^{\text{dCRT}}(X, Y, Z)$ is defined as in (4). We use the same fitting procedures for the GCM test as for spaCRT and dCRT.
- The **negative binomial regression score test**, a standard score test implemented via the `glm.nb()` function in the `MASS` package. Note this test is based on standard normal distribution to compute the p -values.

For negative binomial regression, the estimation of size parameter is a crucial step. GCM, dCRT and spaCRT use the simple method of moments estimator, while the score test uses a more sophisticated estimator that requires iterative computation. The detailed discussion of these dispersion estimation methods and other details of testing methods can be found in Appendix K.1. The comparison of different methods applied is summarized in Table 1. We applied both left- and right-sided variants of each test. All simulations are repeated 10,000 times for accurate Type-I error estimation for small p -value thresholds.

Test	Dispersion estimation	Resampling required	Normality based
GCM test	Precomputed	No	Yes
Score test	Iterative	No	Yes
dCRT	Precomputed	Yes	No
spaCRT	Precomputed	No	No

Table 1: Summary table for testing methods compared.

Simulation results Here, we present a representative selection of simulation results (Figure 2). These results correspond to $r = 0.05$ and $\beta_0 = -5$, and all tests are applied at nominal level $\alpha = 0.01$. Additional results are provided in Appendix K.2. We find from Figure 2a, which displays p -value distributions under the null hypothesis, that the spaCRT and dCRT tests have similar p -value distributions, both of which are close to uniform. Meanwhile, the GCM test behaves too liberally for left-sided tests and too conservatively for right-sided tests, while the score test behaves too conservatively for left-sided tests and too liberally for right-sided tests. These trends are reflected in the Type-I error rates and powers in Figure 2b,c. We remark that the spaCRT and dCRT tests control Type-I error for all settings of γ_0 , though both tests tend to become conservative as X becomes sparser. Furthermore, the spaCRT and dCRT are the most powerful tests among those that have Type-I error control for every parameter setting.

Next, we remark on how the methods' performance is impacted by the problem parameters γ_0 and β_0 . As either X or Y become less sparse (i.e., as γ_0 or β_0 increase),

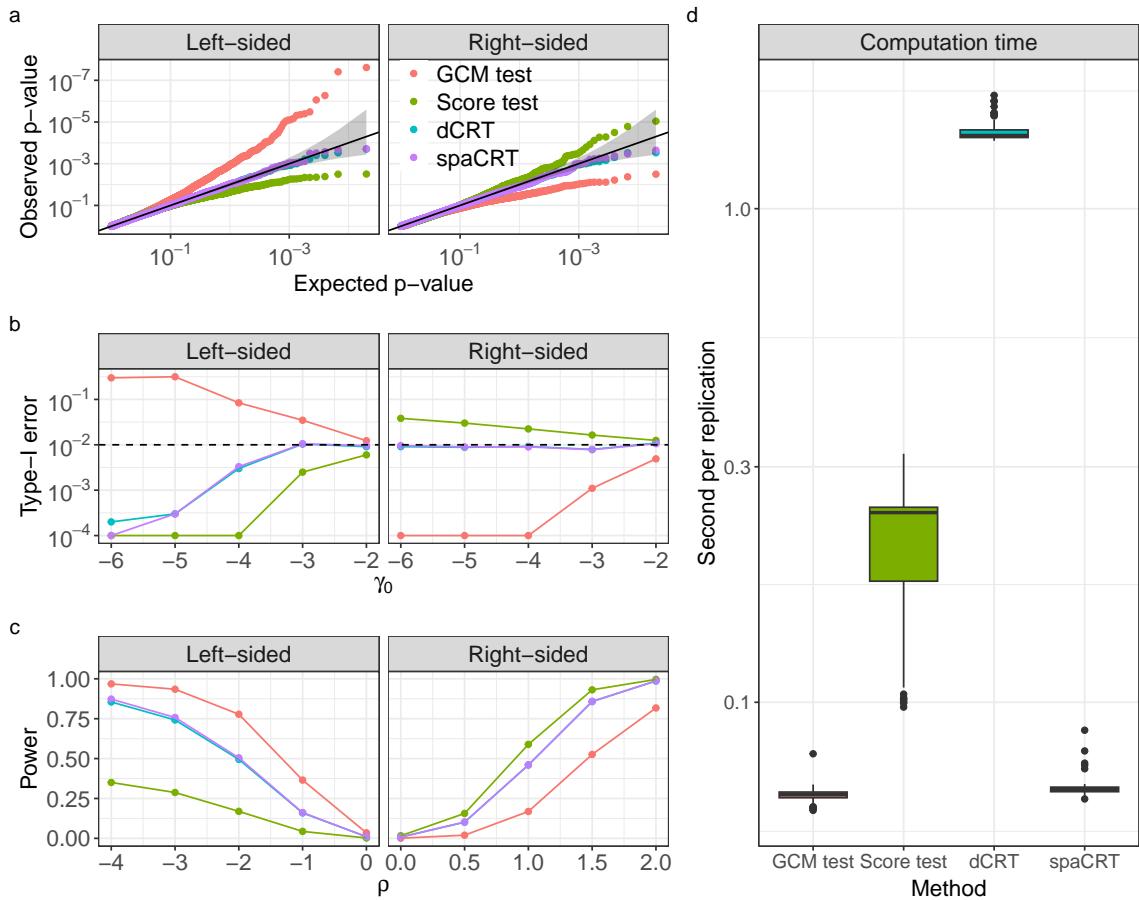


Figure 2: Summary of CRISPR screens simulation results for size parameter $r = 0.05$ and the significance level is set to 0.01. (a) QQ-plots of the p -values obtained under the null hypothesis for $(\gamma_0, \beta_0) = (-3, -5)$. (b) Type-I error rates for $\beta_0 = -5$ as a function of the sparsity of X (γ_0). (c) Power for $(\gamma_0, \beta_0) = (-3, -5)$ as a function of the signal strength (ρ). (d) Computation times by different methods. The boxplot shows the results across different simulation parameters while each point is the averaged results over 10,000 replications.

the p -value distributions, Type-I error rates, and powers for the GCM and score tests improve. This is to be expected, as the test statistics converge more quickly towards the standard normal distribution when the quantities being averaged are less sparse. We defer the discussion on size parameter r to Appendix K.2 where we will show dCRT and spaCRT are more robust to the values of r than the GCM and score tests.

Moreover, Figure 2d displays the computing times for the different methods in the settings considered in Figure 2a,b,c. We see that the spaCRT and GCM test are roughly tied for fastest, the score test is roughly half an order of magnitude slower than these two, while the dCRT is more than an order of magnitude slower.

Even though the CRISPR screens analysis is often concerned about multiple genes and perturbations (i.e. high-multiplicity), our simulation setup already captures the relevant statistical phenomena on the single perturbation-gene pair. To complete our story, we regard the simulation setup as being under a multiple testing regime and

present the number of rejections in Appendix K.2 after applying multiplicity corrections. These results are consistent with the simulations presented in this section.

6.2 GWAS with rare phenotypes and genetic variations

Source of sparsity in GWAS with rare phenotypes: GWAS aims to investigate if genetic variation is associated with a phenotype of interest. There can be two sources of sparsity in such analysis. On one hand, the phenotype of interest (Y) can be rare, i.e. only a small fraction of the population has the phenotype. Typical examples include certain rare diseases which have low prevalence in the population. On the other hand, the sparsity can come from the rare genetic variation (X), i.e., only a small fraction of the population has the genetic variation. Both sources of sparsity have been a major challenge in GWAS (Auer and Lettre, 2015; Zhao et al., 2020; Dey et al., 2017; Turro et al., 2020).

Simulation setup: The goal of identifying genetic variations associated with phenotypes in GWAS can be formulated as a conditional independence testing problem (Sesia, Sabatti, and Candès, 2019):

$$H_0^j : \mathbf{X}_j \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}_{-j} \quad j = 1, \dots, d \quad (37)$$

where $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_d)^\top \in \{0, 1\}^d$ is typically referring to the presence or absence of a specific allele at a given locus and \mathbf{Y} is the phenotype of interest. Often time, a large scale loci are considered simultaneously. In other words, this is a high-dimensional setup. We consider a binary outcome Y , which may be viewed as indicator of the presence of a disease:

$$\mathbf{Y} \mid \mathbf{X} \sim \text{Ber}(\text{expit}(\gamma_0 + \mathbf{X}^\top \boldsymbol{\beta})).$$

Then under this model, the variable selection problem (37) is now equivalent to testing $\beta_j = 0$ or not (Candès et al., 2018). Similar to model (36), γ_0 here controls the sparsity of outcome Y and we consider $\{-3, -2\}$ for *high* and *low* sparsity settings, mimicking the relative frequency of phenotype occurrence.

We consider the genetic variable $\mathbf{X} \in \{0, 1\}^d$ is generated from a *hidden Markov model* (HMM). HMMs have been broadly adopted to describe haplotypes, the sequence of alleles at a series of markers along one chromosome, and offer a good phenomenological description of the dependence between the explanatory variables in GWAS (Scheet and Stephens, 2006; Marchini et al., 2007; Browning and Browning, 2007). A brief introduction of the HMM can be found in Appendix J. We consider two sets of parameters generating HMMs with one setup mimicking the rare genetic variation scenario (high sparsity in X) and the other setup mimicking the common genetic variation scenario (low sparsity in X).

We will consider $d = 500$ and sample size $n = 2000$ in the simulation and leave the other details on parameters used in the simulation to Appendix K.1.

Methodologies compared: We compare the following four methods.

- The **spaCRT** (Algorithm 2), where the parameters of the HMM, the distribution \mathbf{X} , is fitted based on expectation-maximization (EM) algorithm implemented by **fastPhase** software (Scheet and Stephens, 2006) and **fastPhase** has also been a popular method in the recent variable selection literatures (Sesia, Sabatti, and Candès, 2019). Conditional distribution $\mathbf{Y} \mid \mathbf{X}_j$ is fitted based on a modified high-dimensional logistic regression with lasso penalization (Tibshirani, 1996). We refer the details of methods implementation to Appendix L.1.
- The **Knockoffs** (Barber and Candès, 2015; Sesia, Sabatti, and Candès, 2019), where the distribution of \mathbf{X} is fitted in the same way as in spaCRT and knock-off variables are constructed via backford sampling algorithm proposed in Sesia, Sabatti, and Candès (2019). The test statistic is chosen to be the difference of absolute coefficient values between the variable of interest \mathbf{X}_j and its corresponding knockoff variable $\tilde{\mathbf{X}}_j$, which are obtained in the high-dimensional logistic regression with lasso penalization for fitting $\mathbf{Y}|\mathbf{X}, \tilde{\mathbf{X}}$, where $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_d)^\top \in \mathbb{R}^d$.

We also include **dCRT** (Algorithm 1) with resample $M = 5000$ and **GCM test** (Shah and Peters, 2020), with the same fitting procedures as the spaCRT. We apply a both-sided test for all four methods. For spaCRT, dCRT and GCM, the p -values are corrected for multiple testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The FDR is controlled at level 0.1.

Simulation results We present a representative selection of simulation results (Figure 3) with high sparsity in X (i.e. rare genetic variation) and the other results can be found in Appendix L.2. Note that the Knockoff procedure does not provide p -values but just a rejection set. Thus controlling family-wise error rate (FWER) is more challenging for the Knockoff procedure due to the lack of p -values. Therefore we only present the FDR plot (Figure 3a) and power plot (Figure 3b) in this simulation setup. We find from Figure 3a that the spaCRT and dCRT tests have similar FDR, both of which are close to the nominal level. Meanwhile, the GCM test behaves too liberally due to excessive sparsity in X and Y although the degree of inflation is more severe when $\gamma_0 = -3$. The Knockoff method can also control FDR while being conservative when signal is weak. Figure 3b shows GCM, dCRT and spaCRT have similar power whereas Knockoff procedure tends to be less powerful. We believe the comparably low power in Knockoff may be due to sensitivity of Knockoff procedure to the distribution of \mathbf{X} , which may heavily affect the construction of decorrelated knockoff variable $\tilde{\mathbf{X}}$ to observed variable \mathbf{X} . Thus we are reluctant to claim that Knockoff procedure is less powerful and postpone the comprehensive investigation on this phenomenon to future work.

Figure 3c shows the time consumption for different methods. Among all methods, Knockoff is the computationally most efficient method and this hinges on the fact that only one high-dimensional regression fit of $\mathbf{Y}|\mathbf{X}, \tilde{\mathbf{X}}$ is required, where here $\tilde{\mathbf{X}}$ is the constructed knockoff variable. On the other hand, the other three methods involve d regressions for $\mathbf{Y}|\mathbf{X}_{-j}, j \in [d]$ if no further acceleration is applied, not to mention the excessive the resampling required in dCRT. With this being said, we employ a *tower trick* to boost the computation of $\hat{\mathbb{E}}[\mathbf{Y}|\mathbf{X}_{-j}]$ for dCRT, GCM and spaCRT. The implementation details can be found in Appendix L.1. The acceleration

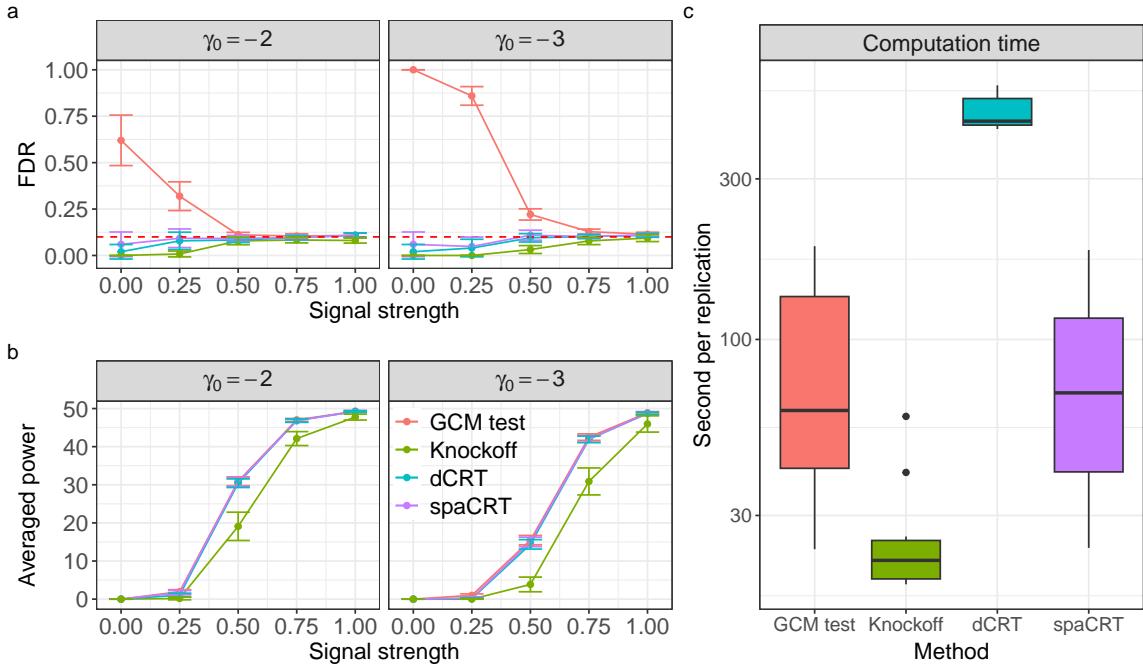


Figure 3: Summary of numerical simulation results for variable selection with high sparsity in data X . All the results are obtained with the regularization parameter $\lambda = \text{lambda.1se}$. (a) FDR for $\gamma_0 = -3$ (high sparsity) and $\gamma_0 = -2$ (low sparsity). (b) Power for the same set of γ_0 . (c) Computation times consumed by different methods across 50 replications and all the parameters.

can push the computation time of spaCRT and GCM to be within a factor of 2 to 3 of the Knockoff procedure while still being much faster than the dCRT. Note we use R package `SNPknock` to generate knockoff variable and the construction procedure is implemented in `Rcpp` for speed. Currently, the other methods are implemented mostly in R so we expect further speedup for these methods when employing more efficient implementation.

7 Real data analysis

In this section, we compare the performance of the spaCRT to those of alternative methods on the analysis of the Gasperini et al. (2019) single-cell CRISPR screen dataset.

7.1 Overview of the data

The Gasperini data contain expression measurements on 13,135 genes and CRISPR perturbations targeting 6,105 regulatory elements in $n = 207,324$ cells. They also contain CRISPR perturbations intended as negative and positive controls. In particular, the data contain 51 non-targeting CRISPR perturbations, which do not target any regulatory element and therefore should have no effect on the expressions of any genes. Furthermore, the data contain 754 CRISPR perturbations targeting genes, rather than

regulatory elements. These serve as positive controls, because they are known a priori to have effects on the expressions of the genes they target. Finally, the data contain measurements on six covariates, including four count-based covariates related to library size, one binary covariate indicating the experimental batch, and one continuous covariate indicating the proportion of reads mapping to mitochondrial genes in each cell.

7.2 Analyses conducted

Hypotheses tested. In order to assess the Type-I error and power of the methods compared, we will use CRISPR perturbations intended as negative and positive controls, respectively. In particular, for Type-I error analysis, we test for association between each of the 51 negative control perturbations and each of 3,000 randomly sampled genes, for a total of $51 \times 3,000 = 153,000$ tests. We subsample the genes to reduce the computational burden of the analysis. To assess the power of each method, we test for association between each of the 754 positive control perturbations targeting genes and the gene they target, for a total of 754 tests.

Methods compared. We compare essentially the same methods as in the numerical simulations (recall Section 6.1). The only difference is that we replace the dCRT with a faster variant implemented in the R package `sceptre` (Barry et al., 2024; Barry et al., 2021), in order to make the analysis computationally feasible. The `sceptre` implementation of the dCRT fits a parametric curve to the resampling distribution of the test statistic based on a smaller number of resamples (a heuristic acceleration that is not theoretically justified). Furthermore, it is implemented in C++ for speed, unlike the other methods we consider, which are implemented in R. We apply left- and right-sided variants of each test on the negative control perturbation-gene pairs. For the positive control pairs, we apply only left-sided tests, since we are testing for a perturbation-induced decrease in gene expression.

7.3 Results

Type-I error. Figure 4 displays QQ plots of the negative control p -values obtained from all four methods. The two tests relying on asymptotic normality, the GCM and score tests, exhibit severe p -value inflation for left- and right-sided tests, respectively. This finding is consistent with our simulation results (Figure 2). On the other hand, the spaCRT and `sceptre` tests control Type-I error well for both left- and right-sided tests. We also report the number of false discoveries on the negative control pairs in Appendix N.1; the message from these results are consistent with that of the single testing results.

Next, we investigate the impact of the problem sparsity on calibration. Following (Barry et al., 2024), we measure sparsity in terms of the *effective sample size* $\sum_{i=1}^n \mathbb{1}(X_i Y_i > 0)$, which measures the number of cells with a given perturbation and nonzero expression of a given gene. Table 2 displays the distribution of effective sample sizes across the negative control perturbation-gene pairs tested, showing that the

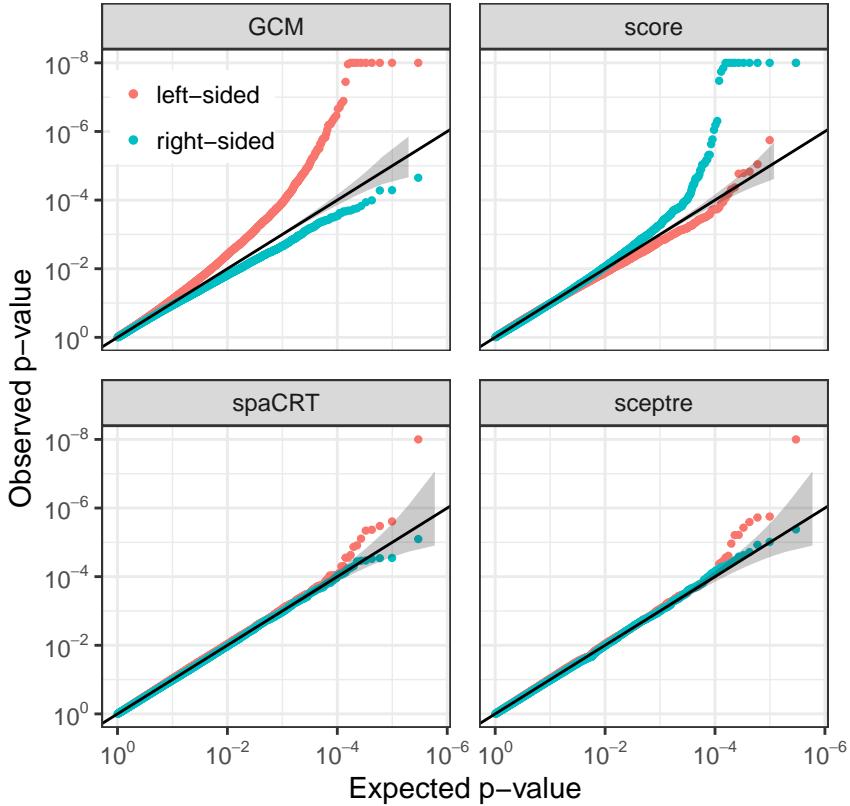


Figure 4: Left- and right-sided p -values for negative control perturbation-gene pairs on the Gasperini data.

effective samples sizes are vastly smaller than the number of cells, $n = 207,324$. Furthermore, Figure 11 stratifies the QQ plots for each method by effective sample size, focusing on those pairs with effective sample size of at most 100. As expected based on our simulation study, we find more severe miscalibration for pairs with lower effective sample sizes, especially for the GCM and score tests, and to a lesser extent for `sceptre` and the spaCRT. We carried out a similar analysis, stratifying the pairs based on the estimated size parameter (Figure 12). In line with our simulation results, we find that the GCM and score tests exhibit more miscalibration for smaller size parameters.

	Min.	1st Qu.	Median	3rd Qu.	Max.
Effective sample size	0	53	204	504	2044

Table 2: Effective sample size in subsampled dataset.

Power. Next, we compare the power of the four methods based on their left-sided p -values on the 754 positive control perturbation-gene pairs (Figure 5). The signal is quite strong in these positive control pairs, as evidenced by small p -values for all four methods. We remark that the spaCRT overcomes the discreteness in the p -values returned by resampling-based methods such as the dCRT, delivering very small p -values in the presence of strong signals. Given the scale of the p -values, we refrain from

making definitive conclusions about the relative power of the methods, but remark only that the spaCRT appears at least as powerful as the alternative methods considered.

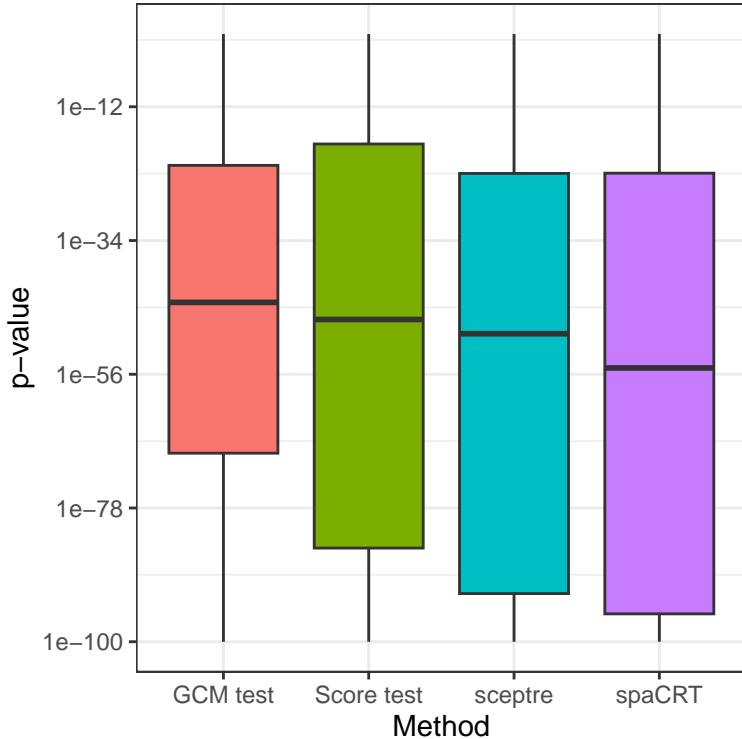


Figure 5: Left-sided p -values computed on the 754 positive control perturbation-gene pairs in the Gasperini et al. (2019) data.

Computation. The excellent statistical properties of the spaCRT do not come at the cost of computational efficiency. Since the `sceptre` software is highly optimized, while the other methods are not, we benchmark our dCRT implementation (used in the numerical simulations) instead of `sceptre`'s for computational efficiency. We use $M = 100,000$ resamples for the dCRT, given the high multiplicity of the problem (recall Section 1.1). To assess running time, we paired two randomly sampled genes with each of the 51 non-targeting perturbations, for a total of 102 perturbation-gene pairs. We find that the spaCRT is roughly as fast as the GCM test, about five times faster than the score test, and about 250 times faster than the dCRT.

Table 3: Computation time per perturbation-gene pair on the Gasperini data. Times are reported in seconds.

Method	Mean	Std dev
GCM test	4.0	1.9
dCRT	1002.1	279.5
spaCRT	4.0	1.9
Score test	19.9	9.3

8 Discussion

In this paper, we introduce the spaCRT, a new conditional independence test enjoying several desirable properties: (1) It is completely resampling-free, which makes it very computationally efficient. (2) It is asymptotically equivalent to the doubly robust dCRT, so it has Type-I error control without requiring the model-X assumption. (3) It has excellent Type-I error control and power in both numerical simulations and real data analysis. The spaCRT is particularly well-suited to data with low signal-to-noise ratio and large scale, such as single-cell CRISPR screen data analysis, where it can significantly accelerate the state-of-the-art `sceptre` software without sacrificing statistical performance.

Despite the attractive properties of the spaCRT, the current work still has several limitations. First, there remains a gap between our theoretical guarantees and the practical performance of the spaCRT and dCRT. We have shown that both of these tests have Type-I error control asymptotically. Except for Theorem 1 where we establish the theoretical gain of odCRT compared to oGCM, we have not theoretically justified why these tests perform so well in finite samples compared with asymptotic tests like the GCM and score tests, when in-sample fitting for $\mu_{n,x}$ and $\mu_{n,y}$ are considered. Second, we have focused on approximating the dCRT in this paper, but the conditional randomization test framework is applicable to general test statistics. It would be interesting to see how the saddlepoint approximation can be extended to other test statistics.

9 Acknowledgments

We acknowledge the Wharton research computing team for their help with our use of the Wharton high-performance computing cluster for the numerical simulations and real data analyses in this paper. This work was partially support by NSF DMS-2113072 and NSF DMS-2310654.

References

- Adamson, Britt et al. (2016). “A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response”. In: *Cell* 167.7, 1867–1882.e21.
- Auer, Paul L and Guillaume Lettre (2015). “Rare variant association studies: considerations, challenges and opportunities”. In: *Genome medicine* 7, pp. 1–11.
- Barber, Rina Foygel and Emmanuel J Candès (2015). “Controlling the false discovery rate via knockoffs”. In: *The Annals of statistics*, pp. 2055–2085.
- Barry, Timothy, Kaishu Mason, Kathryn Roeder, and Eugene Katsevich (2024). “Robust differential expression testing for single-cell CRISPR screens at low multiplicity of infection”. In: *Genome Biology* 25.1, pp. 1–30.
- Barry, Timothy, Xuran Wang, John A. Morris, Kathryn Roeder, and Eugene Katsevich (2021). “SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis”. In: *Genome Biology* 22.

- Benjamini, Yoav and Yosef Hochberg (1995). “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300.
- Bentkus, Vidmantas, Friedrich Götze, and Willem R van Zwet (1997). “An Edgeworth expansion for symmetric statistics”. In: *The Annals of Statistics* 25.2, pp. 851–896.
- Besag, Julian and Peter Clifford (1991). “Sequential Monte Carlo p-values”. In: *Biometrika* 78.2, pp. 301–304.
- Bickel, P. J. (1974). “Edgeworth expansions in nonparametric statistics”. In: *Annals of Statistics* 2.1, pp. 1–20.
- Browning, Sharon R and Brian L Browning (2007). “Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering”. In: *The American Journal of Human Genetics* 81.5, pp. 1084–1097.
- Butler, Ronald (2007). *Saddlepoint approximations with applications*. Cambridge University Press.
- Candès, Emmanuel, Yingying Fan, Lucas Janson, and Jinchi Lv (2018). “Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.3, pp. 551–577.
- Daniels, Henry E. (1954). “Saddlepoint Approximations in Statistics”. In: *The Annals of Mathematical Statistics* 25.4, pp. 631–650.
- Datlinger, Paul et al. (2017). “Pooled CRISPR screening with single-cell transcriptome readout”. In: *Nature Methods* 14.3, pp. 297–301.
- Davidson, James (1994). *Stochastic Limit Theory*. Oxford University Press.
- Davison, Anthony C . and David V. Hinkley (1988). “Saddlepoint Approximations in Resampling Methods”. In: *Biometrika* 75.3, pp. 417–431.
- Dey, Rounak, Ellen M Schmidt, Goncalo R Abecasis, and Seunggeun Lee (2017). “A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS”. In: *The American Journal of Human Genetics* 101.1, pp. 37–49.
- Dixit, Atray et al. (2016). “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens”. In: *Cell* 167, pp. 1853–1866.
- Efron, Bradley (2022). *Exponential Families in Theory and Practice*. Cambridge University Press, pp. 1–250.
- Fischer, Lasse and Aaditya Ramdas (2024a). “Multiple testing with anytime-valid Monte-Carlo p-values”. In: pp. 1–22. arXiv: [2404.15586](https://arxiv.org/abs/2404.15586).
- (2024b). “Sequential Monte-Carlo testing by betting”. In: *arXiv*, pp. 1–33. arXiv: [2401.07365](https://arxiv.org/abs/2401.07365).
- Gandy, Axel (2009). “Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk”. In: *Journal of the American Statistical Association* 104.488, pp. 1504–1511. arXiv: [0612488 \[math\]](https://arxiv.org/abs/0612488).
- Gandy, Axel and Georg Hahn (2014). “MMCTest-A safe algorithm for implementing multiple monte carlo tests”. In: *Scandinavian Journal of Statistics* 41.4, pp. 1083–1101.
- (2016). “A Framework for Monte Carlo based Multiple Testing”. In: *Scandinavian Journal of Statistics* 43.4, pp. 1046–1063.

- Gandy, Axel and Georg Hahn (2017). “QuickMMCTest: quick multiple Monte Carlo testing”. In: *Statistics and Computing* 27.3, pp. 823–832.
- Gasperini, Molly et al. (2019). “A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens”. In: *Cell* 176.1-2, 377–390.e19.
- Ge, Tian, Jianfeng Feng, Derrek P. Hibar, Paul M. Thompson, and Thomas E. Nichols (2012). “Increasing power for voxel-wise genome-wide association studies: The random field theory, least square kernel machines and fast permutation procedures”. In: *NeuroImage* 63.2, pp. 858–873.
- Hall, Peter (2013). *The bootstrap and Edgeworth expansion*. Springer Science & Business Media.
- Huang, Mo et al. (2018). “SAVER: Gene expression recovery for single-cell RNA sequencing”. In: *Nature Methods* 15.7, pp. 539–542.
- Imbens, Guido W and Donald B Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.
- Jaitin, Diego Adhemar et al. (2016). “Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq”. In: *Cell* 167.7, 1883–1896.e15.
- Kolassa, John E. (2006). *Series Approximation Methods in Statistics*. Third Edit. Springer.
- Kuchibhotla, Arun Kumar (2023). “Central Limit Theorems and Approximation Theory: Part II”. In: *arXiv preprint arXiv:2306.14382*.
- Liu, Molei, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas (2022). “Fast and powerful conditional randomization testing via distillation”. In: *Biometrika* 109.2, pp. 277–293.
- Lugannani, Robert and Stephen Rice (1980). “Saddle point approximation for the distribution of the sum of independent random variables”. In: *Advances in Applied Probability* 12.2, pp. 475–490.
- Marchini, Jonathan, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly (2007). “A new multipoint method for genome-wide association studies by imputation of genotypes”. In: *Nature genetics* 39.7, pp. 906–913.
- Niu, Ziang, Abhinav Chakraborty, Oliver Dukes, and Eugene Katsevich (2024). “Reconciling model-X and doubly robust approaches to conditional independence testing”. In: *Annals of Statistics, to appear*.
- Niu, Ziang, Jyotishka Ray Choudhury, and Eugene Katsevich (2024). “The saddlepoint approximation for averages of conditionally independent random variables”. In: *arXiv*.
- Petrov, Valentin V (Apr. 1995). *Oxford Studies In Probability 4: Limit Theorems of Probability Theory Sequences of Independent Random Variables*. Oxford University Press.
- Robinson, J. (1982). “Saddlepoint Approximations for Permutation Tests and Confidence Intervals”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.1, pp. 91–101.
- Scheet, Paul and Matthew Stephens (2006). “A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase”. In: *The American Journal of Human Genetics* 78.4, pp. 629–644.

- Sesia, Matteo, Chiara Sabatti, and Emmanuel J Candès (2019). “Gene hunting with hidden Markov model knockoffs”. In: *Biometrika* 106.1, pp. 1–18.
- Shah, Rajen D. and Jonas Peters (2020). “The Hardness of Conditional Independence Testing and the Generalised Covariance Measure”. In: *Annals of Statistics* 48.3, pp. 1514–1538.
- Svensson, Valentine (2020). “Droplet scRNA-seq is not zero-inflated”. In: *Nature Biotechnology* 38, pp. 142–150.
- Swanson, Jason (2019). *Lecture notes on probability theory*. Tech. rep.
- Tibshirani, Robert (1996). “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1, pp. 267–288.
- Townes, F. William, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry (2019). “Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model”. In: *Genome Biology* 20.1, pp. 1–16.
- Turro, Ernest et al. (2020). “Whole-genome sequencing of patients with rare diseases in a national health system”. In: *Nature* 583.7814, pp. 96–102.
- Wainwright, Martin J (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press.
- Winkler, Anderson M., Gerard R. Ridgway, Gwenaëlle Douaud, Thomas E. Nichols, and Stephen M. Smith (2016). “Faster permutation inference in brain imaging”. In: *NeuroImage* 141, pp. 502–516.
- Zhao, Zhangchen et al. (2020). “UK Biobank whole-exome sequence binary phenotype analysis with robust region-based rare-variant test”. In: *The American Journal of Human Genetics* 106.1, pp. 3–12.

A Probability theory preliminaries

A.1 Single probability space embedding

To better state and understand the conditional convergence result, the following lemma helps to embed all the random variables into one big probability space.

Lemma 1 (Embedding into a single probability space, Lemma 14 in Niu et al., 2024). *Consider a sequence of probability spaces $\{(\mathbb{P}_n, \Omega_n, \mathcal{G}_n), n \geq 1\}$. For each n , let $\{W_{i,n}\}_{i \geq 1}$ be a collection of integrable random variables defined on $(\mathbb{P}_n, \Omega_n, \mathcal{G}_n)$ and let $\mathcal{F}_n \subseteq \mathcal{G}_n$ be a σ -algebra. Then there exists a single probability space $(\tilde{\mathbb{P}}, \tilde{\Omega}, \tilde{\mathcal{G}})$, random variables $\{\tilde{W}_{i,n}\}_{i,n \geq 1}$ on $(\tilde{\mathbb{P}}, \tilde{\Omega}, \tilde{\mathcal{G}})$, and σ -fields $\tilde{\mathcal{F}}_n \subseteq \tilde{\mathcal{G}}$ for $n \geq 1$, such that for each n , the joint distribution of $(\{W_{i,n}\}_{i \geq 1}, \{\mathbb{E}[W_{i,n} | \mathcal{F}_n]\}_{i \geq 1})$ on $(\mathbb{P}_n, \Omega_n, \mathcal{G}_n)$ coincides with that of $(\{\tilde{W}_{i,n}\}_{i \geq 1}, \{\mathbb{E}[\tilde{W}_{i,n} | \tilde{\mathcal{F}}_n]\}_{i \geq 1})$ on $(\tilde{\mathbb{P}}, \tilde{\Omega}, \tilde{\mathcal{G}})$.*

With the above Lemma, we are safe to state any almost sure statement which can be interpreted within one probability space.

A.2 Some facts about natural exponential family

Consider the NEF with probability density function

$$f(x|\theta) = h(x) \exp(\theta x - A(\theta)).$$

Then there is one-to-one correspondence between the moments of the random variable from NEF and the derivative of the log-partition function $A(\theta)$. We summarise the relationship in the following Lemma.

Lemma 2 (Chapter 1.2 in Efron, 2022). *Suppose $X \sim f(x|\theta)$ then the following identities hold:*

1. $\mathbb{E}[X] = A'(\theta);$
2. $\mathbb{E}[X^2] - (\mathbb{E}[X])^2 = A''(\theta);$
3. $\mathbb{E}[(X - \mathbb{E}[X])^3] = A^{(3)}(\theta);$
4. $\mathbb{E}[(X - \mathbb{E}[X])^4] - 3(\mathbb{E}[(X - \mathbb{E}[X])^2])^2 = A^{(4)}(\theta).$

B Preliminaries for saddlepoint approximation

Let W be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ -algebra. Consider the following definitions.

Definition 1 (CSE distribution). *Consider a random variable $\theta \in \mathcal{G}$ such that $\theta \geq 0$ almost surely and a constant $\beta > 0$. We say $W|\mathcal{G}$ is conditionally sub-exponential (CSE) with parameters (θ, β) if, almost surely,*

$$\mathbb{P}[|W| \geq t | \mathcal{G}] \leq \theta \exp(-\beta t), \text{ for all } t > 0.$$

We denote this property via $W|\mathcal{G} \sim \text{CSE}(\theta, \beta)$.

Definition 2 (CCS distribution). Consider a random variable $\nu \in \mathcal{G}$ such that $\nu \geq 0$ almost surely. We say $W|\mathcal{G}$ is conditionally compactly supported (CCS) on $[-\nu, \nu]$ if

$$W \in [-\nu, \nu] \quad \text{almost surely.}$$

We denote this property via $W|\mathcal{G} \sim CCS(\nu)$.

Consider a triangular array $\{W_{in}\}_{1 \leq i \leq n, n \geq 1}$ and σ -algebra \mathcal{F}_n , so that $\mathbb{E}[W_{in}|\mathcal{F}_n] = 0$ for each (i, n) , and $\{W_{in}\}_{1 \leq i \leq n}$ are independent conditionally on \mathcal{F}_n for each n . Now, we impose assumptions on the triangular array $\{W_{in}\}_{1 \leq i \leq n, n \geq 1}$ in terms of Definition 1-2. We assume throughout this section that Assumption 7 or Assumption 8 holds.

Assumption 7 (CSE condition). There exist $\theta_n \in \mathcal{F}_n$ and $\beta > 0$ such that

$$W_{in}|\mathcal{F}_n \sim CSE(\theta_n, \beta) \text{ for all } i, n, \quad \theta_n < \infty \text{ almost surely,} \quad \theta_n = O_{\mathbb{P}}(1). \quad (38)$$

Assumption 8 (CCS condition). There exist $\nu_{in} \in \mathcal{F}_n$ such that

$$W_{in}|\mathcal{F}_n \sim CCS(\nu_{in}), \quad \nu_{in} < \infty \text{ almost surely,} \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n \nu_{in}^4 = O_{\mathbb{P}}(1). \quad (39)$$

Given the sequence $x_n \in \mathcal{F}_n$, we would like to approximate tail probabilities for the average $W_n \equiv \frac{1}{n} \sum_{i=1}^n W_{in}$:

$$\mathbb{P}[W_n \geq x_n|\mathcal{F}_n]$$

We will approximate this tail probability via the saddlepoint approximation. This requires the existence of the conditional cumulant generating function (CGF):

$$K_n(s|\mathcal{F}_n) \equiv \frac{1}{n} \sum_{i=1}^n K_{in}(s|\mathcal{F}_n), \quad K_{in}(s|\mathcal{F}_n) \equiv \log \mathbb{E}[\exp(sW_{in})|\mathcal{F}_n].$$

Here $K_{in}(\cdot | \mathcal{F}_n)$ is the conditional CGF of W_{in} . The very first step in saddlepoint approximation is to find the tilting parameter \hat{s}_n solving the *saddlepoint equation*

$$K'_n(s|\mathcal{F}_n) = x_n. \quad (40)$$

Then the validity of the saddlepoint approximation is formalized in the following two lemmas.

Lemma 3 (Lemma 1 in Niu, Ray Choudhury, and Katsevich, 2024). Suppose Assumption 7 or Assumption 8 holds. Then, there exists a probability-one event \mathcal{A} and an $\varepsilon > 0$ such that, on \mathcal{A} ,

$$K_{in}(s) < \infty \quad \text{for any } s \in (-\varepsilon, \varepsilon) \text{ and for all } i \leq n, \quad n \geq 1. \quad (41)$$

In particular, when Assumption 7 holds, $\varepsilon = \beta/8$ and when Assumption 8 holds, $\varepsilon = 1/8$.

Lemma 4 (Theorem 1 in Niu, Ray Choudhury, and Katsevich, 2024). *Let W_{in} be a triangular array of random variables that are mean-zero and independent for each n , conditionally on \mathcal{F}_n . Suppose either Assumption 7 or Assumption 8 holds, and that*

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[W_{in}^2 | \mathcal{F}_n] = \Omega_{\mathbb{P}}(1). \quad (42)$$

Let $x_n \in \mathcal{F}_n$ be a sequence with $x_n \xrightarrow{\mathbb{P}} 0$ and $\varepsilon > 0$ is defined as in Lemma 3. Then, the saddlepoint equation (40) has a unique and finite solution $\hat{s}_n \in [-\varepsilon/2, \varepsilon/2]$ with probability approaching 1 as $n \rightarrow \infty$. If we define $\lambda_n \equiv \hat{s}_n \sqrt{n K_n''(\hat{s}_n | \mathcal{F}_n)}$ and

$$r_n \equiv \begin{cases} \text{sgn}(\hat{s}_n) \sqrt{2n(\hat{s}_n x_n - K_n(\hat{s}_n | \mathcal{F}_n))} & \text{if } \hat{s}_n x_n - K_n(\hat{s}_n | \mathcal{F}_n) \geq 0; \\ \text{sgn}(\hat{s}_n) & \text{otherwise,} \end{cases}$$

then

$$\mathbb{P} \left[\frac{1}{n} \sum_{i=1}^n W_{in} \geq x_n \mid \mathcal{F}_n \right] = \left(1 - \Phi(r_n) + \phi(r_n) \left\{ \frac{1}{\lambda_n} - \frac{1}{r_n} \right\} \right) (1 + o_{\mathbb{P}}(1)).$$

and

$$\mathbb{P} \left[1 - \Phi(r_n) + \phi(r_n) \left\{ \frac{1}{\lambda_n} - \frac{1}{r_n} \right\} > 0 \right] \rightarrow 1 \text{ as } n \rightarrow \infty.$$

C Some useful lemmas and proofs

Lemma 5 (Lemma 3 in Niu et al., 2024). *Consider two hypothesis tests based on the same test statistic $T_n(X, Y, Z)$ but different critical values:*

$$\phi_n^1(X, Y, Z) \equiv \mathbb{1}(T_n(X, Y, Z) > C_n(X, Y, Z)); \quad \phi_n^2(X, Y, Z) \equiv \mathbb{1}(T_n(X, Y, Z) > z_{1-\alpha}).$$

If the critical value of the first converges in probability to that of the second:

$$C_n(X, Y, Z) \xrightarrow{\mathbb{P}} z_{1-\alpha}$$

and the test statistic does not accumulate near the limiting critical value:

$$\lim_{\delta \rightarrow 0} \limsup_{n \rightarrow \infty} \mathbb{P}_{\mathcal{L}_n}[|T_n(X, Y, Z) - z_{1-\alpha}| \leq \delta] = 0, \quad (43)$$

then the two tests are asymptotically equivalent:

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\mathcal{L}_n}[\phi_n^1(X, Y, Z) = \phi_n^2(X, Y, Z)] = 1.$$

Regularity condition: there exists $\delta > 0$ such that for a sequence of laws \mathcal{L}_n and its estimate $\widehat{\mathcal{L}}_n$, the following assumptions hold:

$$(\widehat{S}_n^{\text{dCRT}})^2 \equiv \frac{1}{n} \sum_{i=1}^n \text{Var}_{\widehat{\mathcal{L}}_n}[X_{in} | Z_{in}] (Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2 = \Omega_{\mathbb{P}}(1); \quad (44)$$

$$\frac{1}{n^{1+\delta/2}} \sum_{i=1}^n |Y_{in} - \widehat{\mu}_{n,y}(Z_{in})|^{2+\delta} \mathbb{E}_{\widehat{\mathcal{L}}_n}[|\widetilde{X}_{in} - \widehat{\mu}_{n,x}(Z_{in})|^{2+\delta} \mid X, Z] = o_{\mathbb{P}}(1); \quad (45)$$

$$\text{Var}_{\widehat{\mathcal{L}}_n}[X_{in} | Z_{in}], (Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2, (Y_{in} - \mu_{n,y}(Z_{in}))^2 < \infty \text{ almost surely.} \quad (46)$$

Lemma 6 (Theorem 9 in Niu et al., 2024). *Let \mathcal{L}_n be a sequence of laws and $\widehat{\mathcal{L}}_n$ be a sequence of estimates. Suppose there exists a sequence of laws \mathcal{L}_n satisfying all the assumptions in **Regularity condition**. Then, the quantile of*

$$T_n^{\text{ndCRT}}(\tilde{X}, X, Y, Z) \equiv \frac{T_n^{\text{dCRT}}(\tilde{X}, X, Y, Z)}{\widehat{S}_n^{\text{dCRT}}} \quad (47)$$

converges to the quantile of the standard normal distribution pointwisely in probability, i.e., for any $p \in (0, 1)$,

$$\mathbb{Q}_p \left[n^{1/2} T_n^{\text{ndCRT}}(\tilde{X}, X, Y, Z) | X, Y, Z \right] \xrightarrow{\mathbb{P}} z_p.$$

Lemma 7 (Corollary 6 in Niu et al., 2024). *Let X_{in} be a triangular array of random variables, such that X_{in} are independent for each n . If for some $\delta > 0$ we have*

$$\frac{1}{n^{1+\delta}} \sum_{i=1}^n \mathbb{E}[|X_{in}|^{1+\delta}] \rightarrow 0, \quad (48)$$

then

$$\frac{1}{n} \sum_{i=1}^n (X_{in} - \mathbb{E}[X_{in}]) \xrightarrow{\mathbb{P}} 0.$$

The condition (48) is satisfied when

$$\sup_{1 \leq i \leq n} \mathbb{E}[|X_{in}|^{1+\delta}] = o(n^\delta).$$

Lemma 8 (Dominance of higher moment). *For any $1 < p < q < \infty$, the following inequality is true almost surely:*

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[|X_{in}|^p | \mathcal{F}_n] \leq \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}[|X_{in}|^q | \mathcal{F}_n] \right)^{p/q}.$$

Lemma 9 (Conditional Hölder inequality, Swanson, 2019, Theorem 6.60). *Let W_1 and W_2 be random variables and let \mathcal{F} be a σ -algebra. If for some $q_1, q_2 \in (1, \infty)$ with $\frac{1}{q_1} + \frac{1}{q_2} = 1$ we have $\mathbb{E}[|W_1|^{q_1}], \mathbb{E}[|W_2|^{q_2}] < \infty$, then*

$$\mathbb{E}[|W_1 W_2| | \mathcal{F}] \leq (\mathbb{E}[|W_1|^{q_1} | \mathcal{F}])^{1/q_1} (\mathbb{E}[|W_2|^{q_2} | \mathcal{F}])^{1/q_2} \quad \text{almost surely.}$$

Lemma 10 (Conditional Jensen inequality, Davidson, 1994, Theorem 10.18). *Let W be a random variable and let ϕ be a convex function, such that W and $\phi(W)$ are integrable. For any σ -algebra \mathcal{F} , we have the inequality*

$$\phi(\mathbb{E}[W | \mathcal{F}]) \leq \mathbb{E}[\phi(W) | \mathcal{F}] \quad \text{almost surely.}$$

Lemma 11 (Conditional Markov's inequality, Davidson, 1994, Theorem 10.17). *Let W be a random variable and let \mathcal{F} be a σ -algebra. If for some $q > 0$, we have $\mathbb{E}[|W|^q] < \infty$, then for any ε we have*

$$\mathbb{P}[|W| \geq \varepsilon | \mathcal{F}] \leq \frac{\mathbb{E}[|W|^q]}{\varepsilon^q} \quad \text{almost surely.}$$

Lemma 12 (Conditional central limit theorem). *Let W_{in} be a triangular array of random variables, such that for each n , W_{in} are independent conditionally on \mathcal{F}_n . Define*

$$S_n^2 \equiv \sum_{i=1}^n \text{Var}[W_{in} \mid \mathcal{F}_n], \quad (49)$$

and assume $\text{Var}[W_{in} \mid \mathcal{F}_n] < \infty$ almost surely for all $i = 1, \dots, n$ and for all $n \in \mathbb{N}$. If for some $\delta > 0$ we have

$$\frac{1}{S_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}[|W_{in} - \mathbb{E}[W_{in} \mid \mathcal{F}_n]|^{2+\delta} \mid \mathcal{F}_n] \xrightarrow{\mathbb{P}} 0, \quad (50)$$

then

$$\frac{1}{S_n} \sum_{i=1}^n (W_{in} - \mathbb{E}[W_{in} \mid \mathcal{F}_n]) \mid \mathcal{F}_n \xrightarrow{d,p} N(0, 1). \quad (51)$$

Lemma 13. *Consider the fixed dimension setup: $(Z_{in}, X_{in}, Y_{in}) = (Z_i, X_i, Y_i)$. Define $\sigma_{\text{dCRT}}^2 = \mathbb{E}[(X_i - \mathbb{E}[X_i \mid Z_i])^2(Y_i - \mathbb{E}[Y_i \mid Z_i])^2]$. Then if $\mathbb{E}[(Y_i - \mathbb{E}[Y_i \mid Z_i])^2] < \infty$ and $\mathbb{P}[X_i, \hat{\mu}_x(Z_i) \in [-S, S]] = 1$ for some $S > 0$, then as long as the following conditions hold:*

$$\begin{aligned} & \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}; \\ & \sigma_{\text{dCRT}}^2 \in (0, \infty); \\ & \frac{1}{n} \sum_{i=1}^n (\mu_x(Z_i) - \hat{\mu}_x(Z_i))^2 = o_{\mathbb{P}}(1); \\ & \frac{1}{n} \sum_{i=1}^n (\mu_x(Z_i) - \hat{\mu}_x(Z_i))^2 = o_{\mathbb{P}}(1); \\ & \left(\frac{1}{n} \sum_{i=1}^n (\mu_x(Z_i) - \hat{\mu}_x(Z_i))^2 \right) \left(\frac{1}{n} \sum_{i=1}^n (\mu_x(Z_i) - \hat{\mu}_x(Z_i))^2 \right) = o_{\mathbb{P}}(1/n), \end{aligned}$$

then we have $\sqrt{n}T_n^{\text{dCRT}} \xrightarrow{\mathbb{P}} N(0, \sigma_{\text{dCRT}}^2)$.

C.1 Proof of Lemma 8

Proof of Lemma 8. By Lemma 9, we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[|X_{in}|^p \mid \mathcal{F}_n] & \leq \frac{1}{n} \left(\sum_{i=1}^n (\mathbb{E}[|X_{in}|^p \mid \mathcal{F}_n])^{q/p} \right)^{p/q} n^{1-p/q} \\ & = \left(\frac{1}{n} \sum_{i=1}^n (\mathbb{E}[|X_{in}|^p \mid \mathcal{F}_n])^{q/p} \right)^{p/q}. \end{aligned}$$

We use Jensen's inequality, Lemma 10, to obtain

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[|X_{in}|^p \mid \mathcal{F}_n] \leq \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}[|X_{in}|^q \mid \mathcal{F}_n] \right)^{p/q}.$$

□

C.2 Proof of Lemma 13

We consider the following decomposition:

$$\begin{aligned}
T_n^{\text{dCRT}} &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu_x(Z_i))(Y_i - \mu_y(Z_i)) + \frac{1}{n} \sum_{i=1}^n (\mu_x(Z_i) - \hat{\mu}_x(Z_i))(Y_i - \mu_y(Z_i)) \\
&\quad + \frac{1}{n} \sum_{i=1}^n (X_i - \mu_x(Z_i))(\mu_y(Z_i) - \hat{\mu}_y(Z_i)) + \frac{1}{n} \sum_{i=1}^n (\mu_x(Z_i) - \hat{\mu}_x(Z_i))(\mu_y(Z_i) - \hat{\mu}_y(Z_i)) \\
&\equiv \frac{1}{n} \sum_{i=1}^n (X_i - \mu_x(Z_i))(Y_i - \mu_y(Z_i)) + \sum_{k=1}^3 B_k.
\end{aligned}$$

We will prove $\sqrt{n}B_k$ converges 0 in probability for any $k = 1, 2, 3$. For B_1 and B_2 , we have

$$\mathbb{P} [\sqrt{n}B_1 > \varepsilon | X, Z] \leq \frac{1}{\varepsilon^2 n} \sum_{i=1}^n (\mu_x(Z_i) - \hat{\mu}_x(Z_i))^2 \mathbb{E}[(Y_i - \mu_y(Z_i))^2 | Z] = o_{\mathbb{P}}(1)$$

and

$$\mathbb{P} [\sqrt{n}B_2 > \varepsilon | X, Z] \leq \frac{1}{\varepsilon^2 n} \sum_{i=1}^n (\mu_y(Z_i) - \hat{\mu}_y(Z_i))^2 \mathbb{E}[(Y_i - \mu_x(Z_i))^2 | Z] = o_{\mathbb{P}}(1).$$

As for B_3 , we know by Cauchy-Schwarz inequality that

$$\sqrt{n}|B_3| \leq \sqrt{n} \sqrt{\frac{1}{n} \sum_{i=1}^n (\mu_x(Z_i) - \hat{\mu}_x(Z_i))^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (\mu_y(Z_i) - \hat{\mu}_y(Z_i))^2} = o_{\mathbb{P}}(1).$$

Therefore, we only need to prove the weak convergence of $\sum_{i=1}^n (X_i - \mu_x(Z_i))(Y_i - \mu_y(Z_i))/\sqrt{n}$. This is true by usual CLT.

D Proof of Theorem 1

D.1 An asymptotic expansion result

First, we need an auxiliary result to prove Theorem 1.

Lemma 14 (Theorem 5.18 in (Petrov, 1995); Theorem 4.1 in (Kuchibhotla, 2023)). *Consider a sequence of independently and identically distributed random variables $W_{in} \in \mathbb{R}$. Suppose $\mathbb{E}[W_{in}] = 0$, $\mathbb{E}[W_{in}^2] = 1$ and $\mathbb{E}[W_{in}^4] < \infty$ for any $n \in \mathbb{N}$. Then there exists a universal constant $C > 0$ such that for all $x \in \mathbb{R}$,*

$$\begin{aligned}
&\left| \mathbb{P} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n W_{in} \leq x \right] - \mathbb{P}[Z \leq x] - \frac{(1-x^2) \exp(-x^2/2) \mathbb{E}[W_{in}^3]}{6\sqrt{2\pi n}} \right| \\
&\leq C \frac{\mathbb{E}[W_{in}^4]}{n} + C \left(\sup_{|t| \geq 1/(12\mathbb{E}[|W_{in}|^3])} |\mathbb{E}[\exp(itW_{in})]| + \frac{1}{2n} \right)^n \frac{n^6}{1+|x|^4}. \tag{52}
\end{aligned}$$

D.2 Proof of Theorem 1

Proof of Theorem 1. This argument for the validity of odCRT is based on the exchangeability of the resampled data and the original data under null hypothesis $H_0 : \mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$. Now we prove the convergence rate of oGCM. In order to prove the result, we will use Lemma 14 to state an asymptotic expansion of CDF of oGCM test statistic (7). We apply $W_{in} = S_{in}$ and $x = z_{1-\alpha}$ in Lemma 14 so that we get a bound as in (8) with explicit r_n defined as

$$r_n = \frac{(1 - z_{1-\alpha}^2) \exp(-z_{1-\alpha}^2/2) \mathbb{E}[S_{in}^3]}{6\sqrt{2\pi n}}.$$

We just need to show that (52) with $W_{in} = S_{in}$ and $x = z_{1-\alpha}$ is of smaller order of r_n . It is sufficient to show the following results:

$$\frac{\mathbb{E}[S_{in}^4]}{nr_n} = o(1) \quad \text{and} \quad \frac{1}{r_n} \left(\sup_{|t| \geq 1/(12\mathbb{E}[|S_{in}|^3])} |\mathbb{E}[\exp(itS_{in})]| + \frac{1}{2n} \right)^n \frac{n^6}{1 + |z_{1-\alpha}|^4} = o(1).$$

To prove these statements, we first show the order of convergence of r_n . Then we can obtain by conditional independence $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$ and Assumption 2 that

$$\begin{aligned} r_n &\sim \frac{1}{n^{1/2}} \frac{\mathbb{E}[\mu_{n,x}(Z_{in})(1 - \mu_{n,x}(Z_{in}))(1 - 2\mu_{n,x}(Z_{in}))\mathbb{E}[(Y_{in} - \mathbb{E}[Y_{in}|Z_{in}])^3|Z_{in}]]}{(\mathbb{E}[\mu_{n,x}(Z_{in})(1 - \mu_{n,x}(Z_{in}))\text{Var}[Y_{in}|Z_{in}]]))^{3/2}} \\ &\sim \frac{1}{n^{1/2}v_n^{1/2}}. \end{aligned}$$

The last equivalence holds by Hölder's inequality. Together with Assumption 3, we know

$$\limsup_{n \rightarrow \infty} \sup_{|t| \geq 1/(12\mathbb{E}[|S_{in}|^3])} |\mathbb{E}[\exp(itS_{in})]| \leq \limsup_{n \rightarrow \infty} \sup_{t \in \mathbb{R}} |\mathbb{E}[\exp(itS_{in})]| \equiv c < 1.$$

Thus we have for any $\gamma > 0$,

$$\left(\sup_{|t| \geq 1/(12\mathbb{E}[|S_{in}|^3])} |\mathbb{E}[\exp(itS_{in})]| + \frac{1}{2n} \right)^n \lesssim (c + 1/2n)^n = o(n^\gamma)$$

so that we prove

$$\frac{1}{r_n} \left(\sup_{|t| \geq 1/(12\mathbb{E}[|S_{in}|^3])} |\mathbb{E}[\exp(itS_{in})]| + \frac{1}{2n} \right)^n \frac{n^6}{1 + |z_{1-\alpha}|^4} = o(1).$$

Then it remains to prove

$$\frac{\mathbb{E}[S_{in}^4]}{nr_n} = o(1). \tag{53}$$

Proof of convergence (53): It suffices to show $\mathbb{E}[S_{in}^4] = o(n^{1/2}/v_n^{1/2})$ by the proved results $r_n \sim 1/(nv_n)^{1/2}$. To see this, by Assumption 2, we can bound

$$\begin{aligned}\mathbb{E}[S_{in}^4] &\leq \frac{\mathbb{E}[\mu_{n,x}(Z_{in})(1 - \mu_{n,x}(Z_{in}))(1 - 3\mu_{n,x}(Z_{in}) + 3\mu_{n,x}^2(Z_{in}))]}{(\mathbb{E}[\mu_{n,x}(Z_{in})(1 - \mu_{n,x}(Z_{in}))])^2} \\ &\quad \times \frac{\sup_z \mathbb{E}[(Y_{in} - \mathbb{E}[Y_{in}|Z_{in}])^4|Z_{in} = z]}{\inf_z \text{Var}^2[Y_{in}|Z_{in} = z]} \\ &\sim \frac{1}{v_n}.\end{aligned}$$

Then by the assumption that $1/v_n = o(n)$, we know $\mathbb{E}[S_{in}^4] = o(n^{1/2}/v_n^{1/2})$. \square

E Proof of results in section 4

We will restate a stronger version of Theorem 3.

Theorem 7 (Stronger version of Theorem 3). *Suppose the assumptions of Theorem 3 hold. Then define the asymptotic test:*

$$\phi_{n,\alpha}^{\text{asy}} \equiv \mathbb{1} \left(\frac{n^{1/2} T_n^{\text{dCRT}}(X, Y, Z)}{\widehat{S}_n^{\text{dCRT}}} > z_{1-\alpha} \right). \quad (54)$$

Then we have

$$\lim_{n \rightarrow \infty} \mathbb{P}_{\mathcal{L}_n} [\phi_{n,\alpha}^{\text{spaCRT}} = \phi_{n,\alpha}^{\text{dCRT}} = \phi_{n,\alpha}^{\text{asy}}] = 1.$$

Consequently, if $n^{1/2} T_n^{\text{dCRT}}(X, Y, Z) / \widehat{S}_n^{\text{dCRT}} \xrightarrow{d} N(0, 1)$, we have $\lim_{n \rightarrow \infty} \mathbb{E}[\phi_{n,\alpha}^{\text{spaCRT}}] = \alpha$.

E.1 Proof of Theorem 2

We have conditional CGF

$$K_{in}(s|\mathcal{F}_n) = A(\widehat{\theta}_{n,x}(Z_{in}) + a_{in}s) - A(\widehat{\theta}_{n,x}(Z_{in})) - a_{in}sA'(\widehat{\theta}_{n,x}(Z_{in})). \quad (55)$$

We will apply Lemma 4 and thus verify the conditions in the lemma. We first verify the variance condition (42).

Verification of (42): Compute $K_{in}''(s \mid \mathcal{F}_n) = a_{in}^2 A''(\widehat{\theta}_{n,x}(Z_{in}) + a_{in}s)$. Then it suffices to guarantee

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[W_{in}^2 | \mathcal{F}_n] = \frac{1}{n} \sum_{i=1}^n K_{in}''(0 \mid \mathcal{F}_n) = \frac{1}{n} \sum_{i=1}^n a_{in}^2 A''(\widehat{\theta}_{n,x}(Z_{in})) = \Omega_{\mathbb{P}}(1).$$

Next we verify assumption 7 and condition (CCS) respectively.

Verification of Assumption 7 with condition (CSE) and (15): We denote the conditional upper tail probability and lower probability respectively as

$$L_{X,\mathcal{F}_n}(a) \equiv \mathbb{P}[X \leq a|\mathcal{F}_n], \quad U_{X,\mathcal{F}_n}(a) \equiv \mathbb{P}[X \geq a|\mathcal{F}_n].$$

Then by the definition of CSE distribution (Definition 1), we can compute

$$\begin{aligned} & \mathbb{P}[W_{in} \geq t|\mathcal{F}_n] \\ &= \mathbb{P}[a_{in}(\tilde{X}_{in} - A'(\hat{\theta}_{n,x}(Z_{in}))) \geq t|\mathcal{F}_n] \\ &= \mathbb{1}(a_{in} > 0)U_{\tilde{X}_{in},\mathcal{F}_n}\left(\frac{t}{a_{in}} + A'(\hat{\theta}_{n,x}(Z_{in}))\right) + \mathbb{1}(a_{in} < 0)L_{\tilde{X}_{in},\mathcal{F}_n}\left(\frac{t}{a_{in}} + A'(\hat{\theta}_{n,x}(Z_{in}))\right) \end{aligned}$$

Then by the definition of natural exponential family, we can write

$$\begin{aligned} & \mathbb{1}(a_{in} > 0)U_{\tilde{X}_{in},\mathcal{F}_n}\left(\frac{t}{a_{in}} + A'(\hat{\theta}_{n,x}(Z_{in}))\right) \\ &= \mathbb{1}(a_{in} > 0) \int_{t/a_{in} + A'(\hat{\theta}_{n,x}(Z_{in}))}^{\infty} \exp(\hat{\theta}_{n,x}(Z_{in})x - A(\hat{\theta}_{n,x}(Z_{in})))h(x)dx \\ &= \mathbb{1}(a_{in} > 0) \int_{t/a_{in} + A'(\hat{\theta}_{n,x}(Z_{in}))}^{\infty} \exp(\hat{\theta}_{n,x}(Z_{in})x + a_{in}x - a_{in}x - A(\hat{\theta}_{n,x}(Z_{in})))h(x)dx \\ &\leq \mathbb{1}(a_{in} > 0) \int_{t/a_{in} + A'(\hat{\theta}_{n,x}(Z_{in}))}^{\infty} \exp((\hat{\theta}_{n,x}(Z_{in}) + a_{in})x - A(\hat{\theta}_{n,x}(Z_{in})))h(x)dx \\ &\quad \times \exp(-t - a_{in}A'(\hat{\theta}_{n,x}(Z_{in}))) \\ &\leq \mathbb{1}(a_{in} > 0) \int_{t/a_{in} + A'(\hat{\theta}_{n,x}(Z_{in}))}^{\infty} \exp((\hat{\theta}_{n,x}(Z_{in}) + a_{in})x - A(a_{in} + \hat{\theta}_{n,x}(Z_{in})))h(x)dx \\ &\quad \times \exp(A(a_{in} + \hat{\theta}_{n,x}(Z_{in})) - A(\hat{\theta}_{n,x}(Z_{in})))\exp(-t - a_{in}A'(\hat{\theta}_{n,x}(Z_{in}))) \\ &\leq \mathbb{1}(a_{in} > 0) \exp(A(\hat{\theta}_{n,x}(Z_{in}) + a_{in}) - A(\hat{\theta}_{n,x}(Z_{in})) - a_{in}A'(\hat{\theta}_{n,x}(Z_{in})))\exp(-t) \\ &\leq \mathbb{1}(a_{in} > 0) \exp(|A(\hat{\theta}_{n,x}(Z_{in}) + a_{in})| + |A(\hat{\theta}_{n,x}(Z_{in}))| + |a_{in}| |A'(\hat{\theta}_{n,x}(Z_{in}))|)\exp(-t) \end{aligned}$$

Similarly, we can derive the upper bound for the lower tail Probability:

$$\begin{aligned} & \mathbb{1}(a_{in} < 0)L_{\tilde{X}_{in},\mathcal{F}_n}\left(\frac{t}{a_{in}} + A'(\hat{\theta}_{n,x}(Z_{in}))\right) \\ &= \mathbb{1}(a_{in} < 0) \int_{-\infty}^{t/a_{in} + A'(\hat{\theta}_{n,x}(Z_{in}))} \exp(\hat{\theta}_{n,x}(Z_{in})x - A(\hat{\theta}_{n,x}(Z_{in})))h(x)dx \\ &= \mathbb{1}(a_{in} < 0) \int_{-\infty}^{t/a_{in} + A'(\hat{\theta}_{n,x}(Z_{in}))} \exp(\hat{\theta}_{n,x}(Z_{in})x + a_{in}x - a_{in}x - A(\hat{\theta}_{n,x}(Z_{in})))h(x)dx \\ &\leq \mathbb{1}(a_{in} < 0) \exp(A(\hat{\theta}_{n,x}(Z_{in}) + a_{in}) - A(\hat{\theta}_{n,x}(Z_{in})) - a_{in}A'(\hat{\theta}_{n,x}(Z_{in})))\exp(-t) \\ &\leq \mathbb{1}(a_{in} < 0) \exp(|A(\hat{\theta}_{n,x}(Z_{in}) + a_{in})| + |A(\hat{\theta}_{n,x}(Z_{in}))| + |a_{in}| |A'(\hat{\theta}_{n,x}(Z_{in}))|)\exp(-t). \end{aligned}$$

Then we have for any $t > 0$,

$$\begin{aligned} & \mathbb{P}[W_{in} \geq t|\mathcal{F}_n] \\ &\leq \exp(|A(\hat{\theta}_{n,x}(Z_{in}) + a_{in})| + |A(\hat{\theta}_{n,x}(Z_{in}))| + |a_{in}| |A'(\hat{\theta}_{n,x}(Z_{in}))|)\exp(-t) \\ &\leq \exp(\sup_i |A(\hat{\theta}_{n,x}(Z_{in}) + a_{in})| + \sup_i |A(\hat{\theta}_{n,x}(Z_{in}))| + \sup_i |a_{in}| |A'(\hat{\theta}_{n,x}(Z_{in}))|)\exp(-t). \end{aligned}$$

Choosing

$$\theta_n = \exp \left(\sup_i |A(\hat{\theta}_{n,x}(Z_{in}) + a_{in})| + \sup_i |A(\hat{\theta}_{n,x}(Z_{in}))| + \sup_i |a_{in}| |A'(\hat{\theta}_{n,x}(Z_{in}))| \right)$$

and $\beta = 1$, we need to verify

$$\theta_n = O_{\mathbb{P}}(1) \text{ and } \theta_n < \infty, \text{ almost surely.}$$

Since by condition (15), we know $\sup_i |a_{in}| \leq \sup_i |Y_{in}| + \sup_i |\hat{\mu}_{n,y}(Z_{in})| < \infty$ almost surely and $|\hat{\theta}_{n,x}(Z_{in})| < \infty$ almost surely, we know $\theta_n < \infty$ almost surely. Now we prove $\theta_n = O_{\mathbb{P}}(1)$. By condition (CSE), we know for any fixed $\delta > 0$, there exists $M(\delta) > 0$ such that

$$\mathbb{P}[\mathcal{S}] \geq 1 - \delta, \text{ where } \mathcal{S} \equiv \left\{ \sup_i |\hat{\theta}_{n,x}(Z_{in})|, \sup_i |a_{in}| \in [0, M(\delta)] \right\}.$$

Then on the event \mathcal{S} , we know

$$\sup_i |A(\hat{\theta}_{n,x}(Z_{in}) + a_{in})| \leq \sup_{x \in [-2M(\delta), 2M(\delta)]} |A(x)|$$

and

$$\sup_i |A'(\hat{\theta}_{n,x}(Z_{in}))| \leq \sup_{x \in [-M(\delta), M(\delta)]} |A'(x)|.$$

Similarly, on the event \mathcal{S} , we have

$$\sup_i |a_{in}| \leq M(\delta), \sup_i |A(\hat{\theta}_{n,x}(Z_{in}))| \leq \sup_{x \in [-2M(\delta), 2M(\delta)]} |A(x)|.$$

Therefore we have

$$\mathbb{P} \left[\theta_n \leq \exp \left(2 \sup_{x \in [-2M(\delta), 2M(\delta)]} |A(x)| + M(\delta) \sup_{x \in [-M(\delta), M(\delta)]} |A'(x)| \right) \right] \geq \mathbb{P}[\mathcal{S}] \geq 1 - \delta.$$

Therefore we have $\theta_n = O_{\mathbb{P}}(1)$. Thus $\varepsilon = \beta/8 = 1/8$ according to Lemma 3.

Verification of Assumption 8 with condition (15) and (CCS): By condition (CCS), we know

$$\mathbb{1}(\tilde{X}_{in} \in [-S, S]) = 1, \text{ almost surely.}$$

This implies that for any $F \in \mathcal{F}_n$, we have

$$\int_F \mathbb{1}(\tilde{X}_{in} \in [-S, S]) d\mathbb{P} = \int_F 1 d\mathbb{P}.$$

Thus we know

$$\mathbb{P}[\tilde{X}_{in} \in [-S, S] | \mathcal{F}_n] = \mathbb{E}[\mathbb{1}(\tilde{X}_{in} \in [-S, S]) | \mathcal{F}_n] = 1, \text{ almost surely.}$$

Thus we have

$$\mathbb{P} \left[a_{in} (\tilde{X}_{in} - \hat{\mu}_{n,x}(Z_{in})) \in [-2|a_{in}|S, 2|a_{in}|S] | \mathcal{F}_n \right] = 1, \text{ almost surely.}$$

Then again by condition (15), we know

$$|a_{in}| \leq |Y_{in}| + |\hat{\mu}_{n,x}(Z_{in})| < \infty, \text{ almost surely.}$$

Moreover, by condition (CCS), we know

$$\frac{1}{n} \sum_{i=1}^n 16S^4 a_{in}^4 = \frac{16S^4}{n} \sum_{i=1}^n (Y_{in} - \hat{\mu}_{n,y}(Z_{in}))^4 = O_{\mathbb{P}}(1).$$

Choosing $\nu_{in} = 2|a_{in}|S$, we complete the proof for CCS distribution. Thus $\varepsilon = 1/8$ according to Lemma 3.

E.2 Proof of Theorem 3 and Theorem 7

Proof of Theorem 3 and Theorem 7. Define the auxiliary test

$$\phi_{n,\alpha}^{\text{aux}} \equiv \mathbb{1} \left(\frac{n^{1/2} T_n^{\text{dCRT}}(X, Y, Z)}{\hat{S}_n^{\text{dCRT}}} > \mathbb{Q}_{1-\alpha(1+M_n)} \left[n^{1/2} T_n^{\text{ndCRT}}(\tilde{X}, X, Y, Z) | X, Y, Z \right] \right)$$

where $T_n^{\text{ndCRT}}(\tilde{X}, X, Y, Z)$ is defined in (47) and M_n is the sequence such that

$$\mathbb{P} \left(T_n^{\text{dCRT}}(\tilde{X}, X, Y, Z) \geq T_n^{\text{dCRT}}(X, Y, Z) | \mathcal{F}_n \right) = p_{\text{spaCRT}}(1 + M_n), \text{ almost surely.}$$

Notice the tests $\phi_{n,\alpha}^{\text{spaCRT}}$ and $\phi_{n,\alpha}^{\text{aux}}$ are equivalent as long as $M_n \in (-1, 1/\alpha - 1)$. Indeed, we have

$$\mathbb{P}[\phi_{n,\alpha}^{\text{spaCRT}} \neq \phi_{n,\alpha}^{\text{aux}}] \leq \mathbb{P}[\hat{S}_n^{\text{dCRT}} = 0] + \mathbb{P}[M_n \notin (-1, 1/\alpha - 1)] \rightarrow 0$$

due to assumption (44) and that $M_n = o_{\mathbb{P}}(1)$. Thus we only need to verify $\phi_{n,\alpha}^{\text{aux}}$ and $\phi_{n,\alpha}^{\text{dCRT}}$ are asymptotically equivalent. Recall the asymptotic test (54). We will prove $\phi_{n,\alpha}^{\text{aux}}, \phi_{n,\alpha}^{\text{asy}}$ are asymptotically equivalent and $\phi_{n,\alpha}^{\text{dCRT}}, \phi_{n,\alpha}^{\text{asy}}$ are asymptotically equivalent, respectively.

Proof of the equivalence of $\phi_{n,\alpha}^{\text{aux}}, \phi_{n,\alpha}^{\text{asy}}$: We apply Lemma 5 with the test statistic $T_n(X, Y, Z)$ to be

$$T_n(X, Y, Z) \equiv \frac{n^{1/2} T_n^{\text{dCRT}}(X, Y, Z)}{\hat{S}_n^{\text{dCRT}}}$$

and cutoff $C_n(X, Y, Z)$ to be

$$C_n(X, Y, Z) \equiv \mathbb{Q}_{1-\alpha(1+M_n)} \left[n^{1/2} T_n^{\text{ndCRT}}(\tilde{X}, X, Y, Z) | X, Y, Z \right].$$

The following lemma characterizes the convergence of $C_n(X, Y, Z)$.

Lemma 15. Suppose the sequence of laws \mathcal{L}_n and its estimate $\widehat{\mathcal{L}}_n$ satisfy all the assumptions in **Regularity condition**. Then for any given $\alpha \in (0, 1)$, we have for any sequence $M_n \in \mathcal{F}_n$ satisfying $M_n = o_{\mathbb{P}}(1)$,

$$\mathbb{Q}_{1-\alpha(1+M_n)} \left[n^{1/2} T_n^{\text{ndCRT}}(\widetilde{X}, X, Y, Z) | X, Y, Z \right] \xrightarrow{\mathbb{P}} z_{1-\alpha}$$

where $T_n^{\text{ndCRT}}(\widetilde{X}, X, Y, Z)$ is defined as in (47).

To apply Lemma 15 (as well as Lemma 6), we will first verify condition (44)-(46) in **Regularity condition** are satisfied under the assumptions of Theorem 3.

Verification of (44): This is true by assumption (16).

Verification of (45): We verify the condition when $\delta = 2$. When condition (CSE) holds, it suffices to prove

$$\frac{1}{n^2} \sum_{i=1}^n \mathbb{E}_{\widehat{\mathcal{L}}_n} [|\widetilde{X}_{in} - \widehat{\mu}_{n,x}(Z_{in})|^4 | X, Z] = o_{\mathbb{P}}(1).$$

By Lemma 2, we know

$$\mathbb{E}_{\widehat{\mathcal{L}}_n} [|\widetilde{X}_{in} - \widehat{\mu}_{n,x}(Z_{in})|^4 | X, Z] = A^{(4)}(\widehat{\theta}_{n,x}(Z_{in})) + 3(A''(\widehat{\theta}_{n,x}(Z_{in})))^2.$$

Since by condition (CSE), $\sup_i |\widehat{\theta}_{n,x}(Z_{in})| = O_{\mathbb{P}}(1)$, we know there exists $\delta > 0$ such that

$$\mathbb{P}[\mathcal{L}] \geq 1 - \delta, \text{ where } \mathcal{L} \equiv \left\{ \sup_i |\widehat{\theta}_{n,x}(Z_{in})| \leq M(\delta) \right\}.$$

Then on the event \mathcal{L} , by the smoothness of function A , we have

$$\begin{aligned} \sup_i |A^{(4)}(\widehat{\theta}_{n,x}(Z_{in}))| &\leq \sup_{x \in [-M(\delta), M(\delta)]} |A^{(4)}(x)| < \infty, \\ \sup_i |A^{(2)}(\widehat{\theta}_{n,x}(Z_{in}))| &\leq \sup_{x \in [-M(\delta), M(\delta)]} |A^{(2)}(x)| < \infty. \end{aligned}$$

Thus we know for any $\delta > 0$,

$$\begin{aligned} \mathbb{P} \left[\sup_i |A^{(4)}(\widehat{\theta}_{n,x}(Z_{in}))| \leq \sup_{x \in [-M(\delta), M(\delta)]} |A^{(4)}(x)| < \infty \right] &\geq \mathbb{P}[\mathcal{L}] \geq 1 - \delta \\ \mathbb{P} \left[\sup_i |A^{(2)}(\widehat{\theta}_{n,x}(Z_{in}))| \leq \sup_{x \in [-M(\delta), M(\delta)]} |A^{(2)}(x)| < \infty \right] &\geq \mathbb{P}[\mathcal{L}] \geq 1 - \delta \end{aligned}$$

This implies

$$\sup_i |A^{(4)}(\widehat{\theta}_{n,x}(Z_{in}))| = O_{\mathbb{P}}(1), \quad \sup_i |A^{(2)}(\widehat{\theta}_{n,x}(Z_{in}))| = O_{\mathbb{P}}(1).$$

Thus we have

$$\begin{aligned} \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}_{\hat{\mathcal{L}}_n} [|\tilde{X}_{in} - \hat{\mu}_{n,x}(Z_{in})|^4 | X, Z] &\leq \frac{1}{n} \left(\sup_i |A^{(4)}(\hat{\theta}_{n,x}(Z_{in}))| + 3 \sup_i (A''(\hat{\theta}_{n,x}(Z_{in})))^2 \right) \\ &= o_{\mathbb{P}}(1). \end{aligned}$$

When condition (CCS) holds, it suffices to prove

$$\frac{1}{n^2} \sum_{i=1}^n (Y_{in} - \hat{\mu}_{n,y}(Z_{in}))^4 = o_{\mathbb{P}}(1).$$

This is true by the condition (CCS).

Verification of (46): $\text{Var}_{\hat{\mathcal{L}}_n}[X_{in}|Z_{in}] = A''(\hat{\theta}(Z_{in})) < \infty$ and $(Y_{in} - \hat{\mu}_{n,y}(Z_{in}))^2 < \infty$ almost surely can be guaranteed respectively by $|\hat{\theta}(Z_{in})| < \infty$ and $|a_{in}| < \infty$ almost surely in assumption (15). As for $(Y_{in} - \mu_{n,y}(Z_{in}))^2 < \infty$, it is true by the integrability of Y_{in} .

Therefore, applying Lemma 15, we have

$$\mathbb{Q}_{1-\alpha(1+M_n)} \left[n^{1/2} T_n^{\text{ndCRT}}(\tilde{X}, X, Y, Z) | X, Y, Z \right] \xrightarrow{\mathbb{P}} z_{1-\alpha}.$$

Moreover, the condition (43) holds for the chosen $T_n(X, Y, Z)$ by condition (20) so that this proves the asymptotic equivalence of $\phi_{n,\alpha}^{\text{aux}}$ and $\phi_{n,\alpha}^{\text{asy}}$.

Proof of the equivalence of $\phi_{n,\alpha}^{\text{aux}}, \phi_{n,\alpha}^{\text{dCRT}}$: In order to prove the asymptotic equivalence between $\phi_{n,\alpha}^{\text{dCRT}}$ and $\phi_{n,\alpha}^{\text{asy}}$, we apply Lemma 5 with the test statistic $T_n(X, Y, Z)$ to be

$$T_n(X, Y, Z) \equiv \frac{n^{1/2} T_n^{\text{dCRT}}(X, Y, Z)}{\hat{S}_n^{\text{dCRT}}}$$

and cutoff $C_n(X, Y, Z)$ to be

$$C_n(X, Y, Z) \equiv \mathbb{Q}_{1-\alpha} \left[n^{1/2} T_n^{\text{ndCRT}}(\tilde{X}, X, Y, Z) | X, Y, Z \right].$$

By Lemma 6, we have proved that under the assumptions in Theorem 3,

$$\mathbb{Q}_{1-\alpha} \left[n^{1/2} T_n^{\text{ndCRT}}(\tilde{X}, X, Y, Z) | X, Y, Z \right] \xrightarrow{\mathbb{P}} z_{1-\alpha}.$$

Similarly, the nonaccumulant assumption (43) has been satisfied by (20) so that we have proved the asymptotic equivalence between $\phi_{n,\alpha}^{\text{dCRT}}$ and $\phi_{n,\alpha}^{\text{asy}}$. \square

E.3 Proof of Corollary 1

Proof of Corollary 1. For any $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}_{H_0}[p_{\text{spaCRT}} \leq \alpha] &= \mathbb{P}_{H_0}[p_{\text{spaCRT}} \leq 0] + \mathbb{P}_{H_0}[p_{\text{spaCRT}} \in (0, \alpha)] \\ &= \mathbb{P}_{H_0}[p_{\text{spaCRT}} \leq 0] + \mathbb{P}_{H_0}[p_{\text{spaCRT}} \in (0, \alpha), p_{\text{dCRT}}/p_{\text{spaCRT}} \leq 1 + \varepsilon] \\ &\quad + \mathbb{P}_{H_0}[p_{\text{spaCRT}} \in (0, \alpha), p_{\text{dCRT}}/p_{\text{spaCRT}} > 1 + \varepsilon] \\ &\leq \mathbb{P}_{H_0}[p_{\text{spaCRT}} \leq 0] + \mathbb{P}_{H_0}[p_{\text{dCRT}}/p_{\text{spaCRT}} > 1 + \varepsilon] \\ &\quad + \mathbb{P}_{H_0}[p_{\text{dCRT}} \leq \alpha(1 + \varepsilon)].\end{aligned}$$

By the asymptotic validity of dCRT, $\lim_{n \rightarrow \infty} \mathbb{P}_{H_0}[p_{\text{dCRT}} \leq \alpha] \leq \alpha$, and conclusion (18) in Theorem 2, we have

$$\lim_{n \rightarrow \infty} \mathbb{P}_{H_0}[p_{\text{spaCRT}} \leq \alpha] \leq \alpha(1 + \varepsilon) + 0 + \lim_{n \rightarrow \infty} \mathbb{P}_{H_0}[p_{\text{spaCRT}} \leq 0].$$

By conclusion (19), we prove

$$\lim_{n \rightarrow \infty} \mathbb{P}_{H_0}[p_{\text{spaCRT}} \leq \alpha] \leq \alpha(1 + \varepsilon).$$

Since $\varepsilon > 0$ is arbitrary, we have $\lim_{n \rightarrow \infty} \mathbb{P}_{H_0}[p_{\text{spaCRT}} \leq \alpha] \leq \alpha$. Therefore, spaCRT is asymptotic valid. \square

E.4 Proof of Lemma 15

Proof of Lemma 15. For any given $\varepsilon \in (0, \min\{1/\alpha - 1, 1\})$, $\eta > 0$, there exists $N(\varepsilon, \eta)$ such that

$$\mathbb{P}[|M_n| > \varepsilon] < \eta, \quad \forall n \geq N(\varepsilon, \eta).$$

We will use T_n^{ndCRT} to denote $T_n^{\text{ndCRT}}(\tilde{X}, X, Y, Z)$. Then consider the $1 - \alpha(1 - \varepsilon)$ and $1 - \alpha(1 + \varepsilon)$ quantiles, we have with probability at least $1 - \eta$, for large enough n , the following is true:

$$\mathbb{Q}_{1-\alpha(1-\varepsilon)}[T_n^{\text{ndCRT}}|X, Y, Z] \geq \mathbb{Q}_{1-\alpha(1+M_n)}[T_n^{\text{ndCRT}}|X, Y, Z] \geq \mathbb{Q}_{1-\alpha(1+\varepsilon)}[T_n^{\text{ndCRT}}|X, Y, Z].$$

Then with probability at least $1 - \eta$, for sufficiently large n , we have

$$|\mathbb{Q}_{1-\alpha(1+M_n)}[T_n^{\text{ndCRT}}|X, Y, Z] - z_{1-\alpha}| \leq A_n + B_n \tag{56}$$

where

$$A_n \equiv |\mathbb{Q}_{1-\alpha(1-\varepsilon)}[T_n^{\text{ndCRT}}|X, Y, Z] - z_{1-\alpha}|, \quad B_n \equiv |\mathbb{Q}_{1-\alpha(1+\varepsilon)}[T_n^{\text{ndCRT}}|X, Y, Z] - z_{1-\alpha}|.$$

Applying Lemma 6, we have

$$\mathbb{Q}_{1-\alpha(1-\varepsilon)}[T_n^{\text{ndCRT}}|X, Y, Z] \xrightarrow{\mathbb{P}} z_{1-\alpha(1-\varepsilon)}, \quad \mathbb{Q}_{1-\alpha(1+\varepsilon)}[T_n^{\text{ndCRT}}|X, Y, Z] \xrightarrow{\mathbb{P}} z_{1-\alpha(1+\varepsilon)}.$$

Thus for the given ε and sufficiently large n , we have with probability at least $1 - \eta$,

$$A_n < \varepsilon + |z_{1-\alpha(1-\varepsilon)} - z_{1-\alpha}|, \quad B_n < \varepsilon + |z_{1-\alpha(1+\varepsilon)} - z_{1-\alpha}|.$$

By the continuity of the quantile function of standard normal distribution, we know there exists a universal constant C_α that only depends on α such that

$$|z_{1-\alpha(1+\varepsilon)} - z_{1-\alpha}| < C_\alpha \varepsilon, \quad |z_{1-\alpha(1-\varepsilon)} - z_{1-\alpha}| < C_\alpha \varepsilon.$$

Then combining (56), we know with probability at least $1 - 2\eta$, for sufficiently large n , we have

$$|\mathbb{Q}_{1-\alpha(1+M_n)}[T_n^{\text{ndCRT}}|X, Y, Z] - z_{1-\alpha}| \leq A_n + B_n < 2C_\alpha \varepsilon + 2\varepsilon.$$

Then since η, ε is arbitrary, we have

$$\mathbb{Q}_{1-\alpha(1+M_n)}[T_n^{\text{ndCRT}}(\tilde{X}, X, Y, Z)|X, Y, Z] = \mathbb{Q}_{1-\alpha(1+M_n)}[T_n^{\text{ndCRT}}|X, Y, Z] \xrightarrow{\mathbb{P}} z_{1-\alpha}.$$

Therefore we complete the proof. \square

F A special case of binary sampling

We will first present a grand result that states the validity of the spaCRT procedure when \mathbf{X} is a binary random variable.

Lemma 16 (Bernoulli sampling). *Suppose \mathbf{X} is a binary variable following the natural exponential family model (3) and Assumption 4 holds. Furthermore, suppose*

$$\frac{1}{n} \sum_{i=1}^n (\mu_{n,y}(Z_{in}) - \hat{\mu}_{n,y}(Z_{in}))^4 = o_{\mathbb{P}}(1), \quad \frac{1}{n} \sum_{i=1}^n (\theta(Z_{in}) - \hat{\theta}_{n,x}(Z_{in}))^2 = o_{\mathbb{P}}(1), \quad (57)$$

and either of the following set of conditions hold:

- **Fourth moment condition:**

$$|\hat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in})| \xrightarrow{a.s.} 0, \quad |\hat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})| \xrightarrow{a.s.} 0, \quad \forall i \in [n], n \in \mathbb{N}_+; \quad (58)$$

$$\mathbb{P}[|\theta(Z_{in})| < \infty] = 1, \quad \sup_n \mathbb{E}_{\mathcal{L}_n}[\mathbf{Y}^4] < \infty. \quad (59)$$

- **Almost sure conditoin:**

$$|\hat{\theta}_{n,x}(Z_{in})| < \infty, |\hat{\mu}_{n,y}(Z_{in})| < \infty \text{ for any } i, n \text{ almost surely}; \quad (60)$$

$$\sup_n \mathbb{E}_{\mathcal{L}_n}[\mathbf{Y}^4] < \infty. \quad (61)$$

- **Bounded condition:**

$$|\hat{\theta}_{n,x}(Z_{in})| < \infty, |\hat{\mu}_{n,y}(Z_{in})| < \infty \text{ for any } i, n \text{ almost surely}; \quad (62)$$

$$\mathbb{P}[\mathbf{Y} \in [-S, S]] = 1 \text{ for some } S > 0. \quad (63)$$

Then if condition (17) is true, the conclusion in Theorem 2 holds and

$$(\widehat{S}_n^{\text{dCRT}})^2 = \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^2(X_{in} - \mu_{n,x}(Z_{in}))^2] + o_{\mathbb{P}}(1). \quad (64)$$

Additionally, if the condition (20) is true, the conclusion in Theorem 3 holds.

Lemma 16 will be used to prove Theorems 4-6.

Remark 5. We can see all these assumptions are very mild conditions. In particular, Assumption 4 is a standard non-degeneracy condition. The conditions (57) and (58) only impose the consistency of estimators without the rate condition requirement. This thus gives much flexibility on estimators $\widehat{\mu}_{n,y}(\cdot)$ and $\widehat{\theta}_{n,x}(\cdot)$. Condition (59) is a standard regularity condition. These conditions, together with Assumption 4, are crucial to show the lower bound (16).

We now divide the proof of Lemma 16 into two parts, depending either the **fourth moment condition** or **bounded condition** is used.

F.1 Proof of Lemma 16 with fourth moment condition

Since \mathbf{X} is a binary variable and follows model (3), we know $\mathbf{X} | \mathbf{Z} \sim \text{Ber}(\text{expit}(\theta(\mathbf{Z})))$. We now verify the conditions in Theorem 2.

Verification of (15): Since $|\theta(Z_{in})| < \infty$ almost surely by condition (59), together with $|\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})| \xrightarrow{a.s.} 0$ in assumption (58), we have

$$|\widehat{\theta}_{n,x}(Z_{in})| \leq |\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})| + |\theta(Z_{in})| < \infty, \text{ almost surely.}$$

We now show $\sup_n \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^2] < \infty$. This implies

$$|Y_{in} - \mu_{n,y}(Z_{in})| < \infty, \forall i \in [n], n \in \mathbb{N}_+ \text{ almost surely.} \quad (65)$$

This is true by using Jensen's inequality (Lemma 10) and Cauchy-Schwarz inequality:

$$\begin{aligned} \sup_n \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^2] &\leq 2 \sup_n (\mathbb{E}[Y_{in}^2] + \mathbb{E}[\mu_{n,y}(Z_{in})^2]) \\ &\leq 2 \sup_n (\mathbb{E}[Y_{in}^2] + \mathbb{E}[Y_{in}^2]) = 4 \sup_n \mathbb{E}[Y_{in}^2] \leq 4 \sqrt{\sup_n \mathbb{E}[Y_{in}^4]} < \infty. \end{aligned}$$

Then by $|\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in})| \xrightarrow{a.s.} 0$ in assumption (58), we have

$$|a_{in}| = |Y_{in} - \widehat{\mu}_{n,y}(Z_{in})| \leq |Y_{in} - \mu_{n,y}(Z_{in})| + |\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in})| < \infty$$

almost surely.

Verification of (16): We can write

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n a_{in}^2 A''(\widehat{\theta}_{n,x}(Z_{in})) &= \frac{1}{n} \sum_{i=1}^n a_{in}^2 A''(\theta(Z_{in})) + \frac{1}{n} \sum_{i=1}^n a_{in} (A''(\widehat{\theta}_{n,x}(Z_{in})) - A''(\theta(Z_{in}))) \\ &\equiv T_1 + T_2.\end{aligned}$$

We will first prove $T_2 = o_{\mathbb{P}}(1)$. To see this, we notice that $A''(x) = \exp(x)/(1+\exp(x))^2$ and it can be easily checked that $A''(x)$ is a lipschitz function with Lipschitz constant 1. Thus we have

$$|T_2| \leq \frac{1}{n} \sum_{i=1}^n a_{in}^2 |\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})| \leq \sqrt{\frac{1}{n} \sum_{i=1}^n a_{in}^4} \sqrt{\frac{1}{n} \sum_{i=1}^n |\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})|^2}.$$

Thus it suffces to show

$$\frac{1}{n} \sum_{i=1}^n a_{in}^4 = O_{\mathbb{P}}(1), \quad \frac{1}{n} \sum_{i=1}^n |\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})|^2 = o_{\mathbb{P}}(1). \quad (66)$$

By assumption (57), it suffcies to show $\frac{1}{n} \sum_{i=1}^n a_{in}^4 = O_{\mathbb{P}}(1)$. In fact, we have

$$\frac{1}{n} \sum_{i=1}^n a_{in}^4 \leq \frac{16}{n} \sum_{i=1}^n (Y_{in} - \mu_{n,y}(Z_{in}))^4 + \frac{16}{n} \sum_{i=1}^n (\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in}))^4.$$

By the convergence of $\widehat{\mu}_{n,y}(Z_{in})$ in assumption (57), it suffices to show $\sup_n \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^4] < \infty$. This is guaranteed by assumption (59) and Jensen's inequality (Lemma 10):

$$\begin{aligned}\sup_n \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^4] &\leq 16 \sup_n (\mathbb{E}[Y_{in}^4] + \mathbb{E}[\mathbb{E}[Y_{in} | Z_{in}]^4]) \\ &\leq 16 \sup_n (\mathbb{E}[Y_{in}^4] + \mathbb{E}[Y_{in}^4]) = 32 \sup_n \mathbb{E}[Y_{in}^4] < \infty.\end{aligned}$$

Therefore, we have proved $T_2 = o_{\mathbb{P}}(1)$ and now we prove $T_1 = \Omega_{\mathbb{P}}(1)$. We first decompose

$$T_1 = \frac{1}{n} \sum_{i=1}^n (Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2 A''(\theta(Z_{in})) \equiv \frac{1}{n} \sum_{i=1}^n (Y_{in} - \mu_{n,y}(Z_{in}))^2 A''(\theta(Z_{in})) + T_3$$

where

$$T_3 \equiv \frac{1}{n} \sum_{i=1}^n \{(Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2 - (Y_{in} - \mu_{n,y}(Z_{in}))^2\} A''(\theta(Z_{in})).$$

Then by the boundedness of A'' , we have

$$\begin{aligned}|T_3| &\leq \frac{1}{n} \sum_{i=1}^n |\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in})| |2Y_{in} - \mu_{n,y}(Z_{in}) - \widehat{\mu}_{n,y}(Z_{in})| \\ &\leq \sqrt{\frac{1}{n} \sum_{i=1}^n (\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in}))^2} \sqrt{\frac{2}{n} \sum_{i=1}^n (Y_{in} - \mu_{n,y}(Z_{in}))^2 + \frac{2}{n} \sum_{i=1}^n a_{in}^2}.\end{aligned}$$

We have shown in above that

$$\frac{1}{n} \sum_{i=1}^n (Y_{in} - \mu_{n,y}(Z_{in}))^2 = O_{\mathbb{P}}(1), \quad \frac{1}{n} \sum_{i=1}^n a_{in}^2 = O_{\mathbb{P}}(1).$$

Then by assumption (57), we have

$$\frac{1}{n} \sum_{i=1}^n (\hat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in}))^2 \leq \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in}))^4} = o_{\mathbb{P}}(1).$$

Thus we have proved $T_3 = o_{\mathbb{P}}(1)$. The final step is to prove

$$\frac{1}{n} \sum_{i=1}^n (Y_{in} - \mu_{n,y}(Z_{in}))^2 A''(\theta(Z_{in})) = \Omega_{\mathbb{P}}(1).$$

We apply weak law of large numbers to triangular arrays to conclude the proof. In particular, we apply Lemma 7 with $\delta = 1$ so we need to verify

$$\sup_n \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^4 (A''(\theta(Z_{in})))^4] < \infty.$$

Since $|A''(x)| \leq 1$ for any $x \in \mathbb{R}$, by assumption (59), we have

$$\begin{aligned} \sup_n \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^4 (A''(\theta(Z_{in})))^2] &\leq \sup_n \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^4] \\ &\leq 32 \sup_n \mathbb{E}[Y_{in}^4] < \infty. \end{aligned}$$

Therefore, applying Lemma 7 and assumption 4 we obtain

$$\frac{1}{n} \sum_{i=1}^n (Y_{in} - \mu_{n,y}(Z_{in}))^2 A''(\theta(Z_{in})) = o_{\mathbb{P}}(1) + \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^2 A''(\theta(Z_{in}))] = \Omega_{\mathbb{P}}(1).$$

This also proves the result (64).

Verification of condition (CCS): Since $\mathbb{P}[\tilde{X}_{in} \in [-1, 1] | \mathcal{F}_n] = 1$ almost surely, it suffices to show

$$\frac{1}{n} \sum_{i=1}^n a_{in}^4 = \frac{1}{n} \sum_{i=1}^n (Y_{in} - \hat{\mu}_{n,y}(Z_{in}))^4 = O_{\mathbb{P}}(1).$$

This has been proved in conclusion (66).

F.2 Proof of Lemma 16 with almost sure condition

Checking the proof with **fourth moment condition**, we know the condition (15) is the only part that differs. However, condition (15) has been directly assumed in condition (60). Therefore we complete the proof.

F.3 Proof of Lemma 16 with bounded condition

Verificaitons of conditions (15) and (CCS) are straightforward. As for condition (16), we just note that

$$\sup_n \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^4] \leq (2S)^4 = 16S^4 < \infty \quad \text{and} \quad \mathbb{E}[Y_{in}^4] \leq S^4 < \infty.$$

Then the proof can go through as that with **fourth moment condition**. Thus we complete the proof.

G Proof of Theorem 4

G.1 Proof of the conclusion in Theorem 2

We need to apply Lemma 16 with **almost sure condition**. In particular, the condition (60) is trivially satisfied by the finiteness of the maximum likelihood estimate $\hat{\beta}, \hat{\theta}$. Also condition (61) is true in the low-dimensional setup. Thus it suffices to prove condition (57). We divide the proof into two steps.

Verificaiton of $\frac{1}{n} \sum_{i=1}^n (\mu_{n,y}(Z_{in}) - \hat{\mu}_{n,y}(Z_{in}))^4 = o_{\mathbb{P}}(1)$ We will prove a stronger result

$$\frac{1}{n} \sum_{i=1}^n (\mu_y(Z_i) - \hat{\mu}_y(Z_i))^4 = O_{\mathbb{P}}(1/n^2). \quad (67)$$

Define the set $C \equiv \{t : |t| \leq C_Z(\|\beta\|_1 + \|\theta\|_1) + 1\}$. Then consider the event

$$\mathcal{C} \equiv \{Z_i^\top \hat{\beta}, Z_i^\top \hat{\theta} \in C, \forall i \in [n]\}.$$

On the event \mathcal{C} , we know

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (\mu_y(Z_i) - \hat{\mu}_y(Z_i))^4 &= \frac{1}{n} \sum_{i=1}^n (A'(Z_i^\top \beta) - A'(Z_i^\top \hat{\beta}))^4 \\ &\leq \frac{1}{n} \sum_{i=1}^n \sup_{t \in C} (A''(t))^4 |Z_i^\top (\beta - \hat{\beta})|^4 \quad (\text{By mean value theorem}) \\ &\leq \sup_{t \in C} (A''(t))^4 \sup_i \|Z_i\|_\infty^4 \|\hat{\beta} - \beta\|_1^4 \quad (\text{By Hölder's inequality.}) \\ &\leq \sup_{t \in C} (A''(t))^4 C_Z^4 \|\hat{\beta} - \beta\|_1^4 \quad (\text{By Assumption 5}) \\ &= O_{\mathbb{P}}(1/n^2). \quad (\text{By condition (21)}) \end{aligned}$$

Now it suffices to prove $\mathbb{P}[\mathcal{C}] \rightarrow 1$. To see this, we compute

$$|Z_i^\top \hat{\beta}| \leq \sup_i \|Z_i\|_\infty \|\hat{\beta}\|_1 \leq C_Z \|\hat{\beta}\|_1 \leq C_Z (\|\beta\|_1 + \|\hat{\beta} - \beta\|_1)$$

and

$$|Z_i^\top \hat{\theta}| \leq \sup_i \|Z_i\|_\infty \|\hat{\theta}\|_1 \leq C_Z \|\hat{\theta}\|_1 \leq C_Z (\|\theta\|_1 + \|\hat{\theta} - \theta\|_1).$$

By condition (21), we have $\|\hat{\beta} - \beta\|_1 = o_{\mathbb{P}}(1)$ and $\|\hat{\theta} - \theta\|_1 = o_{\mathbb{P}}(1)$. Thus $\mathbb{P}[\mathcal{C}] \rightarrow 1$.

Verificaiton of $\frac{1}{n} \sum_{i=1}^n (\theta(Z_{in}) - \widehat{\theta}_{n,x}(Z_{in}))^2 = o_{\mathbb{P}}(1)$ We will also prove a stronger result:

$$\frac{1}{n} \sum_{i=1}^n (\theta(Z_{in}) - \widehat{\theta}_{n,x}(Z_{in}))^2 = \frac{1}{n} \sum_{i=1}^n (Z_i^\top \widehat{\theta} - Z_i^\top \theta)^2 = O_{\mathbb{P}}(1/n).$$

We Hölder's inequality, we have

$$\frac{1}{n} \sum_{i=1}^n (Z_i^\top \widehat{\theta} - Z_i^\top \theta)^2 \leq \frac{1}{n} \sum_{i=1}^n \|Z_i\|_\infty^2 \|\widehat{\theta} - \theta\|_1^2 \leq C_Z^2 \|\widehat{\theta} - \theta\|_1^2 = O_{\mathbb{P}}(1/n).$$

G.2 Proof of the asymptotic validity under null

We have verified the conditions for Lemma 16 to hold in section G.1. Thus we know by conclusion (64)

$$(\widehat{S}_n^{\text{dCRT}})^2 \xrightarrow{\mathbb{P}} \mathbb{E}[(Y_i - \mathbb{E}[Y_i|Z_i])^2(X_i - \mathbb{E}[X_i|Z_i])^2] \equiv \sigma_{\text{dCRT}}^2.$$

Thus by Lemma 13, it is sufficient to prove

$$\sqrt{n} T_n^{\text{dCRT}} \xrightarrow{\mathbb{P}} N(0, \sigma_{\text{dCRT}}^2).$$

It is sufficient to prove the following

$$\frac{1}{n} \sum_{i=1}^n (\mu_x(Z_i) - \widehat{\mu}_x(Z_i))^2 = O_{\mathbb{P}}(1/n) \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n (\mu_y(Z_i) - \widehat{\mu}_y(Z_i))^2 = O_{\mathbb{P}}(1/n).$$

For the first claim, by Cauchy-Schwarz inequality and result (67), we have

$$\frac{1}{n} \sum_{i=1}^n (\mu_x(Z_i) - \widehat{\mu}_x(Z_i))^2 \leq \sqrt{\frac{1}{n} \sum_{i=1}^n (\mu_x(Z_i) - \widehat{\mu}_x(Z_i))^4} = O_{\mathbb{P}}(1/n).$$

For the second claim, we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (\mu_y(Z_i) - \widehat{\mu}_y(Z_i))^2 &= \frac{1}{n} \sum_{i=1}^n (\text{expit}(Z_i^\top \theta) - \text{expit}(Z_i^\top \widehat{\theta}))^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n (Z_i^\top \theta - Z_i^\top \widehat{\theta})^2 \quad (\text{By Lipschitz continuity of expit}) \\ &\leq C_Z \|\widehat{\theta} - \theta\|_1^2 = O_{\mathbb{P}}(1/n). \quad (\text{By Assumption 5 and (21)}) \end{aligned}$$

Therefore we complete the proof.

H Proof of Theorem 5

H.1 Proof of the conclusion in Theorem 2

We first present a lemma which acts as a building block for proving Theorem 5.

Lemma 17. Suppose all the assumptions in Theorem 5 except for (28) hold. Recall λ_n and ν_n as in models (22) and (23). Then if we choose

$$\lambda_n = C_\lambda \sqrt{\log(d)/n} \quad \text{and} \quad \nu_n = C_\nu \sqrt{\log(d)/n}$$

for some universal constants C_λ, C_ν , then conclusion in Theorem 2 hold. Furthermore, the variance convergence (64) holds. If additionalaly, the condition (20) is true, the conclusion in Theorem 3 holds.

The proof of Lemma 17 will be postponed to section H.4.

H.2 Proof of the asymptotic validity under null

By Lemma 17 and Theorem 3, in order to prove the power and validity of spaCRT, it suffices to show the following conditions are satisfied:

$$\frac{n^{1/2}T_n^{\text{dCRT}}(X, Y, Z)}{\widehat{S}_n^{\text{dCRT}}} \xrightarrow{d} N(0, 1).$$

The proof is then divided to two parts:

1. We first show the convergence of the test statistic $n^{1/2}T_n^{\text{dCRT}}(X, Y, Z)$ to a normal distribution;
2. We then show the convergence of the variance $(\widehat{S}_n^{\text{dCRT}})^2$ to the variance of the normal distribution derived.

Then we finish the proof by Slutsky's theorem. In fact, the variance convergence $(\widehat{S}_n^{\text{dCRT}})^2$ has been proved as in Lemma 17. We will dedicate the rest of the proof to show the convergence of $n^{1/2}T_n^{\text{dCRT}}(X, Y, Z)$. We first decompose $n^{1/2}T_n^{\text{dCRT}}(X, Y, Z)$ as follows:

$$n^{1/2}T_n^{\text{dCRT}}(X, Y, Z) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_{in} - \mathbb{E}[X_{in}|Z_{in}])(Y_{in} - \mathbb{E}[Y_{in}|Z_{in}]) + \text{Bias}_1 + \text{Bias}_2 + \text{Bias}_3$$

where

$$\begin{aligned} \text{Bias}_1 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_{in} - \mathbb{E}[X_{in}|Z_{in}])(\mathbb{E}[Y_{in}|Z_{in}] - \widehat{\mu}_{n,y}(Z_{in})) \\ \text{Bias}_2 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbb{E}[X_{in}|Z_{in}] - \widehat{\mu}_{n,x}(Z_{in}))(Y_{in} - \mathbb{E}[Y_{in}|Z_{in}]) \\ \text{Bias}_3 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbb{E}[X_{in}|Z_{in}] - \widehat{\mu}_{n,x}(Z_{in}))(\mathbb{E}[Y_{in}|Z_{in}] - \widehat{\mu}_{n,y}(Z_{in})). \end{aligned}$$

We will now show that these biases will go to 0 in probability.

Lemma 18 (Bias term convergence). Suppose all the assumptions in Theorem 5 hold. Then we have

$$\text{Bias}_1, \text{Bias}_2, \text{Bias}_3 = o_{\mathbb{P}}(1).$$

Now we finish the proof by applying Lemma 12 with $W_{in} = (X_{in} - \mu_{n,x}(Z_{in}))(Y_{in} - \mu_{n,y}(Z_{in}))/\sqrt{n}$, $\mathcal{F}_n = \{\emptyset, \Omega\}$ and $\delta = 2$

$$\frac{\mathbb{E}[(X_{in} - \mu_{n,x}(Z_{in}))^4(Y_{in} - \mu_{n,y}(Z_{in}))^4]}{\mathbb{E}[(X_{in} - \mu_{n,x}(Z_{in}))^2(Y_{in} - \mu_{n,y}(Z_{in}))^2]^2 n} \xrightarrow{\mathbb{P}} 0.$$

converges to 0 in probability. This is true because of the bound

$$\begin{aligned} & \mathbb{E}[(X_{in} - \mu_{n,x}(Z_{in}))^4(Y_{in} - \mu_{n,y}(Z_{in}))^4] \leq \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^4] \\ &= \mathbb{E}[A_y^{(4)}(Z^\top \theta_n) + 3(A_y''(Z^\top \theta_n))^2] \leq \sup_{\|t\|_2 \leq C_Z \sup_n \|\theta\|_n} (A_y^{(4)}(t) + 3(A_y''(t))^2) \end{aligned}$$

and $\inf_n \mathbb{E}[(X_{in} - \mu_{n,x}(Z_{in}))^2(Y_{in} - \mu_{n,y}(Z_{in}))^2] > 0$ by Assumption 4. Then we have

$$\frac{\sum_{i=1}^n (X_{in} - \mu_{n,x}(Z_{in}))(Y_{in} - \mu_{n,y}(Z_{in}))}{\sqrt{n \mathbb{E}[(X_{in} - \mu_{n,x}(Z_{in}))^2(Y_{in} - \mu_{n,y}(Z_{in}))^2]}} \xrightarrow{d} N(0, 1).$$

In other words, by Slutsky's theorem, we have proved

$$\frac{n^{1/2} T_n^{\text{dCRT}}}{\widehat{S}_n^{\text{dCRT}}} \xrightarrow{d} N(0, 1).$$

Then we know (43) is satisfied. Thus by Theorem 7, we know

$$\lim_{n \rightarrow \infty} \mathbb{E}[\phi_{n,\alpha}^{\text{spaCRT}}] = \lim_{n \rightarrow \infty} \mathbb{E}[\phi_{n,\alpha}^{\text{dCRT}}] = \lim_{n \rightarrow \infty} \mathbb{P}[\phi_{n,\alpha}^{\text{asy}}] = \alpha.$$

H.3 A strong consistency result for the lasso estimators

Proof of Lemma 17 hinges on a general consistency results proved in (Wainwright, 2019).

Lemma 19 (A modified version of Corollary 9.26 in (Wainwright, 2019)). *Consider the lasso estimators (22) and (23). Suppose assumptions 5-6 and condition (26) hold. If we choose $\lambda_n = C_\lambda \sqrt{\log(d)/n}$ and $\nu_n = C_\nu \sqrt{\log(d)/n}$ for some universal constants C_λ, C_ν , then for any $\varepsilon > 0$, there exists $N(\varepsilon) \in \mathbb{N}$ such that whenever $n \geq N(\varepsilon)$, we have*

$$\mathbb{P}[\|\widehat{\theta}_n - \theta\|_1 > \varepsilon] \leq 2 \exp(-2n^{1-\delta})$$

and

$$\mathbb{P}[\|\widehat{\beta}_n - \beta_n\|_1 > \varepsilon] \leq 2 \exp(-2n^{1-\delta}).$$

Consequently, $\|\widehat{\theta}_n - \theta\|_1$ and $\|\widehat{\beta}_n - \beta_n\|_1$ converge to 0 almost surely.

The result is a direct consequence of the concentration inequality for the lasso estimators proved in Corollary 9.26 in (Wainwright, 2019). We now give the proof of the lemma.

Proof of Lemma 19. We need a combination of of Corollary 9.26 and Theorem 9.36 in (Wainwright, 2019) to prove Lemma 19.

Lemma 20 (A combination of Corollary 9.26 and Theorem 9.36 in (Wainwright, 2019)). Consider the lasso estimators (22) and (23). Suppose assumptions 5-6 holds. If we choose $\lambda_n = C_\lambda \sqrt{\log(d)/n}$ and $\nu_n = C_\nu \sqrt{\log(d)/n}$ for some universal constants C_λ, C_ν , then we have

$$\mathbb{P}[\|\hat{\theta}_n - \theta_n\|_1 > C_1 s_{\theta_n} \sqrt{\log(d)/n}] \leq 2 \exp(-2 \log(d))$$

and

$$\mathbb{P}[\|\hat{\beta}_n - \beta_n\|_1 > C_2 s_{\beta_n} \sqrt{\log(d)/n}] \leq 2 \exp(-2 \log(d)).$$

To prove Lemma 19, it is sufficient to show for some $\delta \in (0, 1)$,

$$s_{\theta_n} \sqrt{\log(d)/n} = o(1), \quad s_{\beta_n} \sqrt{\log(d)/n} = o(1), \quad \log(d) \sim n^{1-\delta}.$$

These conditions are satisfied by condition (26). \square

H.4 Proof of Lemma 17

We prove the results by applying Lemma 16 combined with the result in Lemma 19. We need to verify conditions (17), (57)-(59).

Verificaiton of (17) Defining $b_{in} \equiv X_{in} - \hat{\mathbb{E}}[X_{in}|Z_{in}]$, we consider the following decomposition:

$$\begin{aligned} T_n^{\text{dCRT}} &= \frac{1}{n} \sum_{i=1}^n b_{in}(Y_{in} - \mathbb{E}[Y_{in}|Z_{in}]) + \frac{1}{n} \sum_{i=1}^n b_{in}(\mathbb{E}[Y_{in}|Z_{in}] - \hat{\mu}_{n,y}(Z_{in})) \\ &\equiv A_n + B_n. \end{aligned}$$

It thus suffices to show $A_n = o_{\mathbb{P}}(1)$ and $B_n = o_{\mathbb{P}}(1)$. We first show $A_n = o_{\mathbb{P}}(1)$ by observing that A_n is just a sample average of conditioally independent random variables on data (X, Z) . Since

$$\begin{aligned} \mathbb{E}[A_n^2|X, Z] &= \frac{1}{n^2} \sum_{i=1}^n b_{in}^2 \mathbb{E}[(Y_{in} - \mathbb{E}[Y_{in}|Z_{in}])^2|X, Z] \\ &\leq \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}[(Y_{in} - \mathbb{E}[Y_{in}|Z_{in}])^2|X, Z] \quad (|b_{in}| \leq 1 \text{ almost surely}) \\ &= \frac{1}{n^2} \sum_{i=1}^n A_y''(Z_{in}^\top \beta_n) \quad (\text{Conditional independence}) \\ &\leq \frac{1}{n} \max_{t \in B} A_y''(t) \quad \text{where } B \equiv \{t \in \mathbb{R} : |t| \leq \sup_n \|\beta_n\|_1 C_Z + 1\}, \\ &\quad (\text{Compactness of } X, Z \text{ and Hölder's inequality}) \end{aligned}$$

then we know $\mathbb{E}[A_n^2] < \max_{t \in B} A_y''(t)$. By conditional Markov's inequality (Lemma 11): for any $\varepsilon > 0$,

$$\mathbb{P}[|A_n| > \varepsilon|X, Z] \leq \frac{\mathbb{E}[A_n^2|X, Z]}{\varepsilon^2} \leq \frac{\max_{t \in B} A_y''(t)}{n \varepsilon^2}$$

almost surely. Taking expectation on both sides, we have $A_n = o_{\mathbb{P}}(1)$. Now we show $B_n = o_{\mathbb{P}}(1)$. We observe

$$|\mathbb{E}[Y_{in}|Z_{in}] - \widehat{\mu}_{n,y}(Z_{in})| = |A'_y(Z_{in}^\top \beta_n) - A'_y(Z_{in}^\top \widehat{\beta}_n)|. \quad (68)$$

On the event

$$E_n \equiv \left\{ Z_{in}^\top \beta_n, Z_{in}^\top \widehat{\beta}_n \in B \text{ for any } i \in [n] \right\}, \quad (69)$$

we know

$$\begin{aligned} |B_n| &\leq \frac{1}{n} \sum_{i=1}^n |b_{in}(\mathbb{E}[Y_{in}|Z_{in}] - \widehat{\mu}_{n,y}(Z_{in}))| && \text{(Triangle inequality)} \\ &\leq \frac{1}{n} \sum_{i=1}^n |(\mathbb{E}[Y_{in}|Z_{in}] - \widehat{\mu}_{n,y}(Z_{in}))| && (|b_{in}| \leq 1 \text{ almost surely}) \\ &= \frac{1}{n} \sum_{i=1}^n |A'_y(Z_{in}^\top \beta_n) - A'_y(Z_{in}^\top \widehat{\beta}_n)| && \text{(By result (68))} \\ &\leq \frac{1}{n} \sum_{i=1}^n \max_{t \in B} A''_y(t) |Z_{in}^\top (\widehat{\beta}_n - \beta_n)| && \text{(Mean value theorem)} \\ &\leq \max_{t \in B} A''_y(t) C_Z \|\widehat{\beta}_n - \beta_n\|_1. && \text{(Hölder's inequality)} \end{aligned}$$

By Lemma 19, we know $\|\widehat{\beta}_n - \beta_n\|_1$ converge to 0 almost surely. Thus it suffices to show $\mathbb{P}[E_n] \rightarrow 0$. In fact, we can prove a stronger result.

Lemma 21. *Suppose the assumptions in Theorem 17 hold. Then*

$$\mathbb{P}[E_n^c \text{ happens infinitely often}] = 0.$$

Proof of Lemma 21 can be found in section H.6. With Lemma 21, we conclude the verificaiton for (17).

Verificaiton of (57) We first show $(1/n) \sum_{i=1}^n (\mu_{n,y}(Z_{in}) - \widehat{\mu}_{n,y}(Z_{in}))^4 = o_{\mathbb{P}}(1)$. We know one the event E_n (defined in (69)),

$$|A'_y(Z_{in}^\top \widehat{\beta}_n)| \leq \max_{t \in B} A''_y(t) (C_Z \|\widehat{\beta}_n - \beta_n\|_1).$$

Thus we know on the event E_n ,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (\mu_{n,y}(Z_{in}) - \widehat{\mu}_{n,y}(Z_{in}))^4 &= \frac{1}{n} \sum_{i=1}^n (A'(Z_{in}^\top \beta_n) - A'(Z_{in}^\top \widehat{\beta}_n))^4 \\ &\leq \sup_{t \in B} (A''(t))^4 \frac{1}{n} \sum_{i=1}^n |Z_{in}^\top (\widehat{\beta}_n - \beta_n)|^4 && \text{(By mean value theorem)} \\ &\leq \sup_{t \in B} (A''(t))^4 C_Z^4 \|\widehat{\beta}_n - \beta_n\|_1^4. && \text{(By Hölder's inequality)} \end{aligned}$$

Thus by Lemma 19, we know $\|\widehat{\beta}_n - \beta_n\|_1 \xrightarrow{\mathbb{P}} 0$ under the given assumption.

Verification of (58) We first show $|\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in})| \rightarrow 0$ almost surely. It suffices to prove these claims on the event E_n by Lemma 21. In fact, on E_n , using again the mean value theorem, we can prove

$$\begin{aligned} |\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in})| &= |A'(Z_{in}^\top \widehat{\beta}_n) - A'(Z_{in}^\top \beta_n)| \\ &\leq \sup_{t \in B} |A''(t)| |Z_{in}^\top (\widehat{\beta}_n - \beta_n)| \\ &\leq \sup_{t \in B} |A''(t)| C_Z \|\widehat{\beta}_n - \beta_n\|_1 \rightarrow 0 \end{aligned}$$

The last convergence holds almost surely by Lemma 19. Now we prove the claim that $|\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})| \rightarrow 0$ almost surely. We can bound, using Hölder's inequality,

$$|\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})| \leq |Z_{in}^\top \widehat{\theta}_n - Z_{in}^\top \theta_n| \leq C_Z \|\widehat{\theta}_n - \theta_n\|_1 \rightarrow 0$$

almost surely by Lemma 19. This concludes the verification of (58).

Verification of (59) We first show $|\theta(Z_{in})| < \infty$ almost surely. This is because of the following bound:

$$|\theta(Z_{in})| = |Z_{in}^\top \theta_n| \leq \|Z_{in}\|_\infty \|\theta_n\|_1 \leq C_Z \sup_n \|\theta_n\|_1 < \infty.$$

Now we prove the claim $\sup_n \mathbb{E}_{\mathcal{L}_n} [\mathbf{Y}^4] < \infty$. By the representation $\mathbb{E}[(\mathbf{Y} - \mathbb{E}[\mathbf{Y} | \mathbf{Z}])^4] = \mathbb{E}[A_y^{(4)}(\mathbf{Z}^\top \theta_n) + 3(A_y''(\mathbf{Z}^\top \theta_n))^2]$, we can bound

$$\begin{aligned} \mathbb{E}[\mathbf{Y}^4] &\leq 16\mathbb{E}[(\mathbf{Y} - \mathbb{E}[\mathbf{Y} | \mathbf{Z}])^4] + 16\mathbb{E}[(\mathbb{E}[\mathbf{Y} | \mathbf{Z}])^4] \\ &\leq 16\mathbb{E}[A_y^{(4)}(\mathbf{Z}^\top \theta_n)] + 48\mathbb{E}[(A_y''(\mathbf{Z}^\top \theta_n))^2] + 16\mathbb{E}[(A_y'(\mathbf{Z}^\top \theta_n))^4]. \end{aligned}$$

It suffices to prove there exists a universal constant C_M such that the following three statements:

$$\max\left\{\sup_n A_y^{(4)}(\mathbf{Z}^\top \theta_n), \sup_n A_y''(\mathbf{Z}^\top \theta_n), \sup_n A_y'(\mathbf{Z}^\top \theta_n)\right\} \leq C_M < \infty.$$

This can be shown by noticing

$$\sup_n A_y^{(4)}(\mathbf{Z}^\top \theta_n) \leq \sup_{t \in B} |A_y^{(4)}(t)|, \quad \sup_n A_y''(\mathbf{Z}^\top \theta_n) \leq \sup_{t \in B} |A_y''(t)|$$

and

$$\sup_n A_y'(\mathbf{Z}^\top \theta_n) \leq \sup_{t \in B} |A_y'(t)|.$$

Thus C_M can be chosen to be $\max\{\sup_{t \in B} |A_y^{(4)}(t)|, \sup_{t \in B} |A_y''(t)|, \sup_{t \in B} |A_y'(t)|\}$. This concludes the verification of (59).

H.5 Proof of Lemma 18

For Bias_1 , by conditional Markov's inequality (Lemma 11) we have

$$\mathbb{P}[\text{Bias}_1 \geq \varepsilon | Y, Z] \leq \frac{\mathbb{E}[\text{Bias}_1^2 | Y, Z]}{\varepsilon^2}.$$

Then we have

$$\begin{aligned}\mathbb{E}[\text{Bias}_1^2 | Y, Z] &= \frac{1}{n} \sum_{i=1}^n (\text{expit})'(Z_{in}^\top \theta_n)(\mu_{n,y}(Z_{in}) - \hat{\mu}_{n,y}(Z_{in}))^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n (\mu_{n,y}(Z_{in}) - \hat{\mu}_{n,y}(Z_{in}))^2. \quad (\text{Derivative of expit}(\cdot) \text{ is bounded.})\end{aligned}$$

Similarly, for Bias_2 , we have

$$\begin{aligned}\mathbb{P}[\text{Bias}_2 \geq \varepsilon | X, Z] &\leq \frac{\mathbb{E}[\text{Bias}_2^2 | X, Z]}{\varepsilon^2} \\ &\leq \frac{1}{\varepsilon^2 n} \sum_{i=1}^n \mathbb{E}[(Y_{in} - \mathbb{E}[Y_{in} | Z_{in}])^2 | Z_{in}] (\mu_{n,x}(Z_{in}) - \hat{\mu}_{n,x}(Z_{in}))^2 \\ &= \frac{1}{\varepsilon^2 n} \sum_{i=1}^n A_y''(Z_{in}^\top \theta_n) (\mu_{n,x}(Z_{in}) - \hat{\mu}_{n,x}(Z_{in}))^2 \\ &\leq \frac{\sup_{\|t\|_2 \leq C_Z \|\theta_n\|_1} A_y''(t)}{\varepsilon^2 n} \sum_{i=1}^n (\mu_{n,x}(Z_{in}) - \hat{\mu}_{n,x}(Z_{in}))^2. \\ &\quad (\text{By Hölder's inequality})\end{aligned}$$

For Bias_3 , we use Cauchy-Schwarz inequality to bound

$$|\text{Bias}_3| \leq \sqrt{n} \sqrt{\frac{1}{n} \sum_{i=1}^n (\mu_{n,x}(Z_{in}) - \hat{\mu}_{n,x}(Z_{in}))^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (\mu_{n,y}(Z_{in}) - \hat{\mu}_{n,y}(Z_{in}))^2}.$$

Thus combining above observations, it suffice to prove

$$\frac{1}{n} \sum_{i=1}^n (\mu_{n,x}(Z_{in}) - \hat{\mu}_{n,x}(Z_{in}))^2 = o_p(1), \quad \frac{1}{n} \sum_{i=1}^n (\mu_{n,y}(Z_{in}) - \hat{\mu}_{n,y}(Z_{in}))^2 = o_p(1) \quad (70)$$

and

$$\left(\frac{1}{n} \sum_{i=1}^n (\mu_{n,x}(Z_{in}) - \hat{\mu}_{n,x}(Z_{in}))^2 \right) \cdot \left(\frac{1}{n} \sum_{i=1}^n (\mu_{n,y}(Z_{in}) - \hat{\mu}_{n,y}(Z_{in}))^2 \right) = o_p(1/n). \quad (71)$$

We compute

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n (\mu_{n,x}(Z_{in}) - \hat{\mu}_{n,x}(Z_{in}))^2 &\leq \frac{1}{n} \sum_{i=1}^n (Z_{in}^\top \hat{\theta}_n - Z_{in}^\top \theta_n)^2 \quad (\text{Lipschitz property}) \\ &\leq C_Z \|\hat{\theta}_n - \theta_n\|_1^2 \quad (\text{Hölder's inequality})\end{aligned}$$

Then by Lemma 19, we know $\|\hat{\theta}_n - \theta_n\|_1^2 = O_p(s_{\theta_n}^2 \log(d)/n)$. Thus we have

$$\frac{1}{n} \sum_{i=1}^n (\mu_{n,x}(Z_{in}) - \hat{\mu}_{n,x}(Z_{in}))^2 = O_p(s_{\theta_n}^2 \log(d)/n).$$

We can compute

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n (\mu_{n,y}(Z_{in}) - \hat{\mu}_{n,y}(Z_{in}))^2 \\ &= \frac{1}{n} \sum_{i=1}^n (\mathbb{E}[Y_{in}|Z_{in}] - A'_y(Z_{in}^\top \hat{\beta}_n) \text{expit}(Z_{in}^\top \hat{\theta}_n) - A'_y(Z_{in}^\top \hat{\beta}_n)(1 - \text{expit}(Z_{in}^\top \hat{\theta}_n)))^2 \\ &\leq \frac{2}{n} \sum_{i=1}^n (A'_y(Z_{in}^\top \beta_n) - A'_y(Z_{in}^\top \hat{\beta}_n))^2 + \frac{2}{n} \sum_{i=1}^n (A'_y(Z_{in}^\top \beta_n) - A'_y(Z_{in}^\top \hat{\beta}_n))^2 \\ &\leq \frac{2}{n} \sum_{i=1}^n \sup_{t \in B} (A''_y(t))^2 (Z_{in}^\top \hat{\beta}_n - Z_{in}^\top \beta_n)^2 + \frac{2}{n} \sum_{i=1}^n \sup_{t \in B} (A''_y(t))^2 (Z_{in}^\top \hat{\beta}_n - Z_{in}^\top \beta_n)^2 \\ &= O(\|\hat{\beta}_n - \beta_n\|_1^2) \\ &= O_p(s_{\beta_n}^2 \log(d)/n). \end{aligned}$$

where the last inequality is due to mean value theorem and B is defined as in the event (69). Therefore combining the above results and conditions (26) and (28), we know claims (70) and (71) hold so that we finish the proof.

H.6 Proof of Lemma 21

Proof of Lemma 21. First consider $Z_{in}^\top \beta_n$. By Hölder's inequality, we have

$$|Z_{in}^\top \beta_n| \leq C_Z \sup_n \|\beta_n\|_1 \leq C_Z \sup_n \|\beta_n\|_n + 1.$$

Now we consider $Z_{in}^\top \hat{\beta}_n$. Similarly, we have

$$|Z_{in}^\top \hat{\beta}_n| \leq C_Z \|\hat{\beta}_n - \beta_n\|_1 + C_Z \|\beta_n\|_1.$$

Since $\|\hat{\beta}_n - \beta_n\|_1 \rightarrow 0$ almost surely by Lemma 19, we know

$$\begin{aligned} \mathbb{P}[Z_{in}^\top \hat{\beta}_n \in B, \forall i \in [n]] &\geq \mathbb{P}[C_Z \|\hat{\beta}_n - \beta_n\|_1 + C_Z \|\beta_n\|_1 \leq C_Z \sup_n \|\beta_n\|_1 + 1] \\ &\geq \mathbb{P}[C_Z \|\hat{\beta}_n - \beta_n\|_1 \leq 1] \geq \mathbb{P}[C_Z \|\hat{\beta}_n - \beta_n\|_1 \leq 1]. \end{aligned}$$

Thus we know

$$\mathbb{P}[E_n^c] \leq 2\mathbb{P}[\|\hat{\beta}_n - \beta_n\|_1 > 1/C_Z].$$

By Lemma 19, we know there exists $N(1/C_Z)$ such that for any $n \geq N(1/C_Z)$,

$$\mathbb{P}[\|\hat{\beta}_n - \beta_n\|_1 > 1/C_Z] \leq \exp(-2n^{1-\delta}).$$

Then we have

$$\sum_{n=1}^{\infty} \mathbb{P}[E_n^c] = \sum_{n=1}^{\infty} \mathbb{P}[\|\hat{\beta}_n - \beta_n\|_1 > 1/C_Z] \leq N(1/C_Z) + 2 \sum_{n=1}^{\infty} \exp(-2n^{1-\delta}) < \infty.$$

This concludes the proof. \square

I Proof of Theorem 6

We divide the proof to two parts: proof of the approximation accuracy conclusion in Theorem 2 and proof of the asymptotic validity of spaCRT under the null hypothesis. They will be presented in section I.2 and section I.3 respectively. We first present a general consistency result on the KRR estimator which will be used in both proofs.

I.1 A general upper bound on the KRR estimator

Before proceeding to the proof, we present a key lemma which states the finite-sample bound on the in-sample mean square error of the KRR estimator. The statement is a combination of Lemma 30 and 31 in Shah and Peters (2020).

Lemma 22 (KRR finite-sample bound). *Let $Z_1, \dots, Z_n \in \mathbb{R}^d$ be independently and identically distributed observations and suppose for any $j \neq i$,*

$$X_i = f(Z_i) + \varepsilon_i \quad \text{where} \quad \text{Var}[\varepsilon_i] < \sigma^2 \text{ and } \text{Cov}[\varepsilon_i, \varepsilon_i] = 0$$

for $f \in \mathcal{H}$ for some RKHS $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. Consider the kernel matrix $K \in \mathbb{R}^{n \times n}$ have $ijth$ $K_{ij} = k(Z_i, Z_j)/n$ and denote the eigenvalue in the eigen-expansion (29) of kernel k as $\kappa_1 \geq \kappa_2 \geq \dots \geq \kappa_n \geq 0$, then, for any $\lambda > 0$, the regularization estimator takes the form

$$\hat{f}_\lambda \equiv K(K + \lambda I)^{-1}X \quad \text{where} \quad X = (X_1, \dots, X_n)^\top \in \mathbb{R}^n,$$

and satisfies

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[(f(Z_i) - \hat{f}_\lambda(Z_i))^2 | Z_1, \dots, Z_n] \leq \frac{\sigma^2}{\lambda} \frac{1}{n} \sum_{i=1}^n \min\{\kappa_i/4, \lambda\} + \|f\|_{\mathcal{H}}^2 \frac{\lambda}{4}.$$

A direct corollary for Lemma 22 is the convergence of mean square error as a consequence of conditional Markov's inequality (Lemma 11).

Corollary 2 (Convergence of KRR estimator). *Suppose the conditions in Lemma 22 hold. Then if condition (35) holds, we have*

$$\frac{1}{n} \sum_{i=1}^n (f(Z_i) - \hat{f}_\lambda(Z_i))^2 = o_{\mathbb{P}}(1).$$

I.2 Proof of the conclusion in Theorem 2

We will verify the conditions in the version of Lemma 16 with **bounded condition**. In fact, condition (63) is clearly satisfied by the boundedness of \mathbf{Y} (condition (32)) and we now prove the other conditions in the lemma.

Verification of condition (57) We first show, by Corollary 2 and condition (34),

$$\frac{1}{n} \sum_{i=1}^n (\mu_y(Z_i) - \hat{\mu}_y(Z_i))^4 \leq \frac{1}{n} \sum_{i=1}^n (\mu_y(Z_i) - \hat{\mu}_y(Z_i))^2 (S + \sup_{z \in \mathbb{R}^d} |\hat{\mu}_y(z)|)^2 = o_{\mathbb{P}}(1).$$

Then we show

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (\theta(Z_{in}) - \hat{\theta}_{n,x}(Z_{in}))^2 &= \frac{1}{n} \sum_{i=1}^n (Z_i^\top \hat{\theta} - Z_i^\top \theta)^2 \\ &\leq \sup_i \|Z_i\|_\infty \|\hat{\theta} - \theta\|_1 \quad (\text{By H\"older's inequality}) \\ &\leq C_Z \|\hat{\theta} - \theta\|_1 \quad (\text{By Assumption 5}) \\ &= o_{\mathbb{P}}(1). \quad (\text{By condition (33)}) \end{aligned}$$

This completes the verification of condition (57).

Verificaiton of condition (62) The $\hat{\theta}_{n,x}(Z_{in}) = Z_i^\top \hat{\theta}$ is finite almost surely by the definition of maximum likelihood estimator. We will show that $|\hat{\mu}_y(Z_i)| < \infty$ almost surely for any i . This is obvious by noticing the KRR estimators:

$$\hat{\mu}_y(Z_i) = K_{Z_i} (K + \lambda_n I)^{-1} Y, \quad \text{where } K_{Z_i} = (K_{1i}, \dots, K_{ni}).$$

Then by the finiteness and positivity of λ_n and the positive semidefiniteness of the kernel matrix K , we know the claim is true.

Verificaiton of condition (17) We prove a stronger result:

$$\sqrt{n} T_n^{\text{dCRT}} \xrightarrow{d} N(0, \sigma_{\text{dCRT}}^2) \quad \text{where } \sigma_{\text{dCRT}}^2 = \mathbb{E}[(X_i - \mu_x(Z_i))^2(Y_i - \mu_y(Z_i))^2]. \quad (72)$$

By Lemma 13, it suffices to show the following

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (\mu_x(Z_i) - \hat{\mu}_x(Z_i))^2 &= o_{\mathbb{P}}(1), \quad \frac{1}{n} \sum_{i=1}^n (\mu_y(Z_i) - \hat{\mu}_y(Z_i))^2 = o_{\mathbb{P}}(1) \\ \left(\frac{1}{n} \sum_{i=1}^n (\mu_x(Z_i) - \hat{\mu}_x(Z_i))^2 \right) \left(\frac{1}{n} \sum_{i=1}^n (\mu_y(Z_i) - \hat{\mu}_y(Z_i))^2 \right) &= o_{\mathbb{P}}(1/n), \end{aligned}$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu_x(Z_i))(Y_i - \mu_y(Z_i)) \xrightarrow{d} N(0, \sigma_{\text{dCRT}}^2).$$

The last claim is justfied by a application of central limit Theorem under the existence of the second moment. We now focus on proving the other claims. For the first claim, we have

$$\frac{1}{n} \sum_{i=1}^n (\mu_x(Z_i) - \hat{\mu}_x(Z_i))^2 \leq \sup_i \|Z_i\|_\infty \|\hat{\theta} - \theta\|_1 \leq C_Z \|\hat{\theta} - \theta\|_1 = o_{\mathbb{P}}(1).$$

For the second claim, we know it is true by Corollary 2. Next, we prove the third claim. This is obvious by the above arugments and condition (33).

I.3 Proof of the asymptotic validity under null

By Theorem 7, we just need to show that

$$\frac{\sqrt{n}T_n^{\text{dCRT}}}{\widehat{S}_n^{\text{dCRT}}} \xrightarrow{d} N(0, 1)$$

Note that this statement implies condition (43) holds. We have proved the conditions required in Lemma 16 in section I.2 so we know by the conclusion (64) that

$$(\widehat{S}_n^{\text{dCRT}})^2 \xrightarrow{\mathbb{P}} \sigma_{\text{dCRT}}^2 \equiv \mathbb{E}[(X_i - \mu_x(Z_i))^2(Y_i - \mu_y(Z_i))^2].$$

We have proved the conclusion (72) so by Slutsky's theorem, the desired claim is true.

J A preliminary introduction to HMM

J.1 Generating genetic variable from multinomial HMM

Consider matrix $X \in \mathbb{R}^{n \times p}$ with i.i.d. rows following the law of $\mathbf{X} \in \{0, 1\}^d$. Each variable of \mathbf{X} , \mathbf{X}_j , can be thought as a copy inherited from either paternal side or maternal side. Multinomial HMM (mHMM) is a mathematical model to model \mathbf{X}_j because of its probabilistic structure mimicking the hereditary nature. Before introducing the mHMM, let us first define a multinomial Markov chain (mMC) as follows:

Definition 3 (multinomial Makrov chain). *We say a random variable $\mathbf{U} \in \{0, 1, \dots, K\}^d$ follows a multinomial Marko Chain distribution, $\mathbf{U} \sim mMC(q, Q)$, if*

$$\mathbf{U}_1 \sim q \quad \text{and} \quad \mathbf{U}_j | \mathbf{U}_{j-1} \sim Q(\cdot | \mathbf{U}_{j-1}),$$

where q is some multinomial distribution supported on $\{0, 1, \dots, K\}$ and $Q(\cdot | \cdot) \in \mathbb{R}^{K \times K}$ is a transition matrix so that $Q(\cdot | \mathbf{U}_{j-1} = u)$ is a transition distribution given the observation \mathbf{U}_{j-1} is u .

Then we consider the following definition for mHMM:

Definition 4 (mHMM). *Consider a p -dimensional random variable $\mathbf{X} \in \{0, 1\}^d$. We say $\mathbf{X} \sim mHMM(\mathbf{U}(q, Q), e)$ is an observation from a multinomial hidden Markov model with a latent multinomial Markov chain $\mathbf{U} \equiv (\mathbf{U}_1, \dots, \mathbf{U}_p)^\top \sim mMC(q, Q)$ and a multinomial emission distribution $\mathbf{X}_j | \mathbf{U}_j \sim e$. In particular, we denote $\mathbb{P}[\mathbf{X}_j = x_j | \mathbf{U}_j = u_j] \equiv e(x_j | u_j)$.*

In particular, we assume $\mathbf{X} \equiv (\mathbf{X}_1, \dots, \mathbf{X}_p)^\top$ is generated from a mHMM:

$$\mathbf{X} \sim mHMM(q, \mathbf{U}(Q, e)).$$

The graphical illustration of the latent DGP for observed genetic variable $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)^\top$ is shown in Figure 6.

The following identities can be easily obtained:

$$\mathbb{P}[\mathbf{X}_j, \mathbf{U}_{j+1} | \mathbf{U}_j] = \mathbb{P}[\mathbf{X}_j | \mathbf{U}_j] \mathbb{P}[\mathbf{U}_{j+1} | \mathbf{U}_j]; \tag{CI}$$

$$\mathbb{P}[\mathbf{X}_{j-1}, \mathbf{U}_j | \mathbf{U}_{1:(j-1)}, \mathbf{X}_{1:(j-2)}] = \mathbb{P}[\mathbf{X}_{j-1}, \mathbf{U}_j | \mathbf{U}_{j-1}], \tag{Markov}$$

where we use $\mathbf{U}_{1:j}$ to denote $(\mathbf{U}_1, \dots, \mathbf{U}_j)^\top \in \mathbb{R}^j$ and same notation is applied to $\mathbf{X}_{1:j}$.

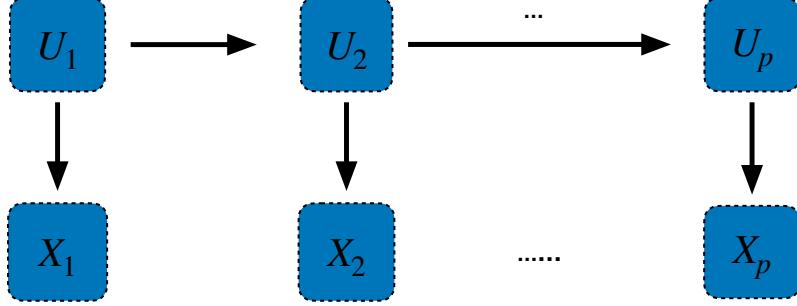


Figure 6: Graphical illustration of a multinomial hidden Markov model. Dashed rounded rectangles: unobserved variables.

J.2 Computing conditional distribution-related quantities

Define $x_{-j} \equiv (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_p)^\top \in \mathbb{R}^{p-1}$. After the joint distribution \mathbf{X} has been estimated from fitting the mHMMs, we can compute the conditional distribution $\mathbf{X}_j | \mathbf{X}_{-j}$ for any $j \in [p]$ with the estimated parameters. We will dedicate this section to discussing how the conditional expectation, conditional cumulant generating function (CCGF) and its derivatives, required by the spaCRT method, can be easily computed. We first boil these quantities down to the conditional probability $\mathbb{P}[\mathbf{X}_j = x_j | \mathbf{X}_{-j} = x_{-j}]$. For the ease of notation, we now define

$$p(x_j, x_{-j}) \equiv \mathbb{P}[\mathbf{X}_j = x_j | \mathbf{X}_{-j} = x_{-j}].$$

1. **Conditional expectation:** For any given $x_{-j} \in \mathbb{R}^{p-1}$, we can compute

$$\mathbb{E}[\mathbf{X}_j | \mathbf{X}_{-j} = x_{-j}] = p(1, x_{-j}) + 2 \cdot p(2, x_{-j}).$$

2. **Function value and derivatives of CCGF:** spaCRT method requires the knowledge of the CCGF function and its derivatives up to second order. Defining $D(t, x_{-j}) \equiv \mathbb{E}[\exp(t\mathbf{X}_j) | \mathbf{X}_{-j} = x_{-j}]$, we can compute

$$D(t, x_{-j}) = p(0, x_{-j}) + \exp(t)p(1, x_{-j}) + \exp(2t)p(2, x_{-j}).$$

Thus we can compute the CCGF value:

$$K(t, x_{-j}) \equiv \log \mathbb{E}[\exp(t\mathbf{X}_j) | \mathbf{X}_{-j} = x_{-j}] = \log(D(t, x_{-j})) \quad \forall t \in \mathbb{R}.$$

Then we compute the first derivative of CCGF:

$$\nabla_t K(t, x_{-j}) = \frac{\exp(t)p(1, x_{-j}) + 2\exp(2t)p(2, x_{-j})}{D(t, x_{-j})}$$

and the second derivative of CCGF:

$$\nabla_t^2 K(t, x_{-j}) = \frac{\exp(t)p(1, x_{-j}) + 4\exp(2t)p(2, x_{-j})}{D(t, x_{-j})} - [\nabla_t K(t, x_{-j})]^2.$$

Thus from the above computation, we know it is sufficient to compute the conditional probability $p(x_j, x_{-j})$. In fact, the following Proposition shows that it can be computed iteratively. To state the Proposition, we need to introduce necessary notation. For $u, \bar{u} \in \{0, 1, \dots, K\}$ and $x \in \{0, 1\}^d$, we define

$$A_j(u, x) \equiv \mathbb{P}[\mathbf{X}_{1:(j-1)} = x_{1:(j-1)}, \mathbf{U}_j = u], \quad A_1(u, x) \equiv \mathbb{P}[\mathbf{U}_1 = u]; \quad (73)$$

$$B_j(u, x) \equiv \mathbb{P}[\mathbf{X}_{(j+1):p} = x_{(j+1):p} | \mathbf{U}_j = u], \quad B_p(u, x) \equiv 1. \quad (74)$$

Proposition 1 (Iterative computation of $p(x_j, x_{-j})$). *Suppose $\mathbf{X} \sim mHMM(\mathbf{U}(q, Q), e)$. Then we have*

$$p(x_j, x_{-j}) = \frac{\sum_{\bar{u} \in \{0, 1, \dots, K\}} e(x_j | \bar{u}) \cdot A_j(\bar{u}, x) \cdot B_j(\bar{u}, x)}{\sum_{\bar{u} \in \{0, 1, \dots, K\}} A_j(\bar{u}, x) \cdot B_j(\bar{u}, x)}. \quad (75)$$

In particular, A_j and B_j can be computed in the following recursive manner: for any $u \in \{0, 1, \dots, K\}$ and $x \in \{0, 1\}^d$,

$$\begin{aligned} A_j(z, x) &= \sum_{\bar{u} \in \{0, 1, \dots, K\}} A_{j-1}(\bar{u}, x) \cdot r(x_{j-1} | \bar{u}) \cdot Q(u | \bar{u}); \\ B_j(u, x) &= \sum_{\bar{u} \in \{0, 1, \dots, K\}} B_{j+1}(\bar{u}, x) \cdot r(x_{j+1} | \bar{u}) \cdot Q(\bar{u} | u). \end{aligned}$$

The proof can be found in Appendix J.3.

J.3 Proof of Proposition 1

Proof of conclusion (75) First, let us consider the following marginalization:

$$\begin{aligned} &\mathbb{P}[\mathbf{X}_j = x_j | \mathbf{X}_{-j} = x_{-j}] \\ &= \sum_{\bar{u} \in \{0, 1, \dots, K\}} \mathbb{P}[\mathbf{X}_j = x_j | \mathbf{X}_{-j} = x_{-j}, \mathbf{U}_j = \bar{u}] \mathbb{P}[\mathbf{U}_j = \bar{u} | \mathbf{X}_{-j} = x_{-j}] \\ &= \sum_{\bar{u} \in \{0, 1, \dots, K\}} \mathbb{P}[\mathbf{X}_j = x_j | \mathbf{U}_j = \bar{u}] \mathbb{P}[\mathbf{U}_j = \bar{u} | \mathbf{X}_{-j} = x_{-j}] \quad (\text{Markov property}) \\ &= \sum_{\bar{u} \in \{0, 1, \dots, K\}} e(x_j | \bar{u}) \mathbb{P}[\mathbf{U}_j = \bar{u} | \mathbf{X}_{-j} = x_{-j}]. \end{aligned}$$

Now we compute the conditioanl probability $\mathbb{P}[\mathbf{U}_j = u | \mathbf{X}_{-j} = x_{-j}]$. In particular, we consider the following decomposition

$$\mathbb{P}[\mathbf{U}_j = u | \mathbf{X}_{-j} = x_{-j}] = \frac{\mathbb{P}[\mathbf{U}_j = u, \mathbf{X}_{-j} = x_{-j}]}{\mathbb{P}[\mathbf{X}_{-j} = x_{-j}]} = \frac{\mathbb{P}[\mathbf{U}_j = u, \mathbf{X}_{-j} = x_{-j}]}{\sum_{\bar{u} \in \{0, 1, \dots, K\}} \mathbb{P}[\mathbf{U}_j = \bar{u}, \mathbf{X}_{-j} = x_{-j}]}.$$

Thus we have

$$\mathbb{P}[\mathbf{X}_j = x_j | \mathbf{X}_{-j} = x_{-j}] = \frac{\sum_{\bar{u} \in \{0, 1, \dots, K\}} e(x_j | \bar{u}) \mathbb{P}[\mathbf{U}_j = \bar{u}, \mathbf{X}_{-j} = x_{-j}]}{\sum_{\bar{u} \in \{0, 1, \dots, K\}} \mathbb{P}[\mathbf{U}_j = \bar{u}, \mathbf{X}_{-j} = x_{-j}]}$$

so that we only need to compute the probability $\mathbb{P}[\mathbf{U}_j = u, \mathbf{X}_{-j} = x_{-j}]$. Further, we can do the following calculation

$$\begin{aligned}\mathbb{P}[\mathbf{U}_j = u, \mathbf{X}_{-j} = x_{-j}] &= \mathbb{P}[\mathbf{X}_1 = x_1, \dots, \mathbf{X}_{j-1} = x_{j-1}, \mathbf{U}_j = u] \\ &\quad \times \mathbb{P}[\mathbf{X}_{j+1} = x_{j+1}, \dots, \mathbf{X}_p = x_p | \mathbf{X}_1 = x_1, \dots, \mathbf{X}_{j-1} = x_{j-1}, \mathbf{U}_j = u] \\ &= \mathbb{P}[\mathbf{X}_1 = x_1, \dots, \mathbf{X}_{j-1} = x_{j-1}, \mathbf{U}_j = u] \\ &\quad \times \mathbb{P}[\mathbf{X}_{j+1} = x_{j+1}, \dots, \mathbf{X}_p = x_p | \mathbf{U}_j = u] \quad (\text{Markov property}) \\ &= A_j(u, x) \cdot B_j(u, x).\end{aligned}$$

The detailed derivation of $A_j(u, x)$ and $B_j(u, x)$ will be present in the next two sections.

Computing $A_j(u, x)$ using forward algorithm We use prove-by-induction to derive $A_j(u, x)$ and the induction is on index j . Since $A_1(u, x) = \mathbb{P}[\mathbf{U}_1 = u]$ and we can compute:

$$\begin{aligned}A_2(u, x) &= \mathbb{P}[\mathbf{X}_1 = x_1, \mathbf{U}_2 = u] \\ &= \sum_{\bar{u} \in \{0, 1, \dots, K\}} \mathbb{P}[\mathbf{U}_1 = \bar{u}] \mathbb{P}[\mathbf{X}_1 = x_1, \mathbf{Z}_2 = z | \mathbf{U}_1 = \bar{u}] \\ &= \sum_{\bar{u} \in \{0, 1, \dots, K\}} \mathbb{P}[\mathbf{U}_1 = \bar{u}] \mathbb{P}[\mathbf{X}_1 = x_1 | \mathbf{U}_1 = \bar{u}] \mathbb{P}[\mathbf{U}_2 = u | \mathbf{U}_1 = \bar{u}] \quad (\text{CI property}) \\ &= \sum_{\bar{u} \in \{0, 1, \dots, K\}} A_1(\bar{u}, x) \cdot e(x_1 | \bar{u}) \cdot Q(u | \bar{u}). \quad (\text{Definition (73)})\end{aligned}$$

Then we can easily show that

$$\begin{aligned}A_j(u, x) &= \mathbb{P}[\mathbf{X}_1 = x_1, \dots, \mathbf{X}_{j-1} = x_{j-1}, \mathbf{U}_j = u] \\ &= \sum_{\bar{u} \in \{0, 1, \dots, K\}} \mathbb{P}[\mathbf{X}_1 = x_1, \dots, \mathbf{X}_{j-2} = x_{j-2}, \mathbf{U}_{j-1} = \bar{u}] \\ &\quad \times \mathbb{P}[\mathbf{X}_{j-1} = x_{j-1}, \mathbf{Z}_j = z | \mathbf{X}_1 = x_1, \dots, \mathbf{X}_{j-2} = x_{j-2}, \mathbf{U}_{j-1} = \bar{u}] \\ &= \sum_{\bar{u} \in \{0, 1, \dots, K\}} \mathbb{P}[\mathbf{X}_1 = x_1, \dots, \mathbf{X}_{j-2} = x_{j-2}, \mathbf{U}_{j-1} = \bar{u}] \\ &\quad \times \mathbb{P}[\mathbf{X}_{j-1} = x_{j-1}, \mathbf{U}_j = u | \mathbf{U}_{j-1} = \bar{u}] \quad (\text{Markov property}) \\ &= \sum_{\bar{u} \in \{0, 1, \dots, K\}} A_{j-1}(\bar{u}, x) \cdot e(x_{j-1} | \bar{u}) \cdot Q(u | \bar{u}). \quad (\text{Definition (73)})\end{aligned}$$

With iterative computation, one can obtain the probability $A_j(u, x)$ for any $j \in [d]$ and $u \in \{0, 1, \dots, K\}$ and $x \in \{0, 1\}^d$.

Computing $B_j(u)$ using backward algorithm Using $B_p(u, x) = 1$ for any u , we can compute

$$\begin{aligned}
B_{p-1}(u, x) &= \mathbb{P}[\mathbf{X}_p = x_p | \mathbf{U}_{p-1} = u] \\
&= \sum_{\bar{u} \in \{0, 1, \dots, K\}} \mathbb{P}[\mathbf{X}_p = x_p, \mathbf{U}_p = \bar{u} | \mathbf{U}_{p-1} = u] \\
&= \sum_{\bar{u} \in \{0, 1, \dots, K\}} \mathbb{P}[\mathbf{X}_p = x_p | \mathbf{U}_p = \bar{u}, \mathbf{U}_{p-1} = u] \mathbb{P}[\mathbf{U}_p = \bar{u} | \mathbf{U}_{p-1} = u] \\
&= \sum_{\bar{u} \in \{0, 1, \dots, K\}} \mathbb{P}[\mathbf{X}_p = x_p | \mathbf{U}_p = \bar{u}] \mathbb{P}[\mathbf{U}_p = \bar{u} | \mathbf{U}_{p-1} = u] \quad (\text{Markov property}) \\
&= \sum_{\bar{u} \in \{0, 1, \dots, K\}} B_p(\bar{u}) \cdot r(x_p | \bar{u}) \cdot s(\bar{u} | u). \quad (\text{Definition (74)})
\end{aligned}$$

By induction, we can compute

$$\begin{aligned}
B_j(u, x) &= \mathbb{P}[\mathbf{X}_{j+1} = x_{j+1}, \dots, \mathbf{X}_p = x_p | \mathbf{U}_j = u] \\
&= \sum_{\bar{u} \in \{0, 1, \dots, K\}} \mathbb{P}[\mathbf{X}_{j+1} = x_{j+1}, \dots, \mathbf{X}_p = x_p, \mathbf{U}_{j+1} = \bar{u} | \mathbf{U}_j = u] \\
&= \sum_{\bar{u} \in \{0, 1, \dots, K\}} \mathbb{P}[\mathbf{X}_{j+2} = x_{j+2}, \dots, \mathbf{X}_p = x_p | \mathbf{U}_{j+1} = \bar{u}, \mathbf{U}_j = u, \mathbf{X}_{j+1} = x_{j+1}] \\
&\quad \times \mathbb{P}[\mathbf{X}_{j+1} = x_{j+1}, \mathbf{Z}_{j+1} = \bar{u} | \mathbf{U}_j = u] \\
&= \sum_{\bar{u} \in \{0, 1, \dots, K\}} \mathbb{P}[\mathbf{X}_{j+2} = x_{j+2}, \dots, \mathbf{X}_p = x_p | \mathbf{U}_{j+1} = \bar{u}] \quad (\text{Markov property}) \\
&\quad \times \mathbb{P}[\mathbf{X}_{j+1} = x_{j+1} | \mathbf{U}_{j+1} = \bar{u}] \cdot \mathbb{P}[\mathbf{U}_{j+1} = \bar{u} | \mathbf{U}_j = u] \quad (\text{CI property}) \\
&= \sum_{\bar{u} \in \{0, 1, \dots, K\}} B_{j+1}(\bar{u}, x) \cdot e(x_{j+1} | \bar{u}) \cdot Q(\bar{u} | u). \quad (\text{Definition (74)})
\end{aligned}$$

With iterative computation, one can obtain the probability $B_j(u, x)$ for any $j \in [d]$ and $u \in \{0, 1, \dots, K\}$ and $x \in \{0, 1\}^d$.

Final form: combining $A_j(u, x)$ and $B_j(u, x)$ Now we compute the probability $\mathbb{P}[\mathbf{U}_j = u_j, \mathbf{X}_{-j} = x_{-j}]$ using the derivation in the above sections. In particular, fixing a set of value x_{-j} , we compute

$$\mathbb{P}[\mathbf{Z}_j = z, \mathbf{X}_{-j} = x_{-j}] = \frac{A_j(z, x) \cdot B_j(z, x)}{\sum_{\bar{z} \in \{0, 1, \dots, K\}} A_j(\bar{z}, x) \cdot B_j(\bar{z}, x)} \quad \forall j \in [d],$$

where $A_j(u, x), B_j(u, x)$ can be computed iteratively as before.

K Additional simulation details in Section 6.1

K.1 Parameters and methods implementation in Section 6.1

Parameters used in Section 6.1 We adopt the parameter settings displayed in Table 4. Note that the bolded values of -5 for γ_0 and β_0 are the default parameter

values. Instead of testing all combinations of these two parameters, we vary one of them while fixing the other to -5. Furthermore, note that our choices of ρ differ based on whether we are carrying out left- or right-sided tests.

γ_0	β_0	ρ (left-sided)	ρ (right-sided)	r	n
-6	-6	-4	0	0.05	5000
-5	-5	-3	0.5	1	
-4	-4	-2	1	10	
-3	-3	-1	1.5		
-2	-2	0	2		

Table 4: Simulation parameter choices.

Methods details in Section 6.1

- The **spaCRT** (Algorithm 2), where $\mathbf{X} | \mathbf{Z}$ is fit based on a logistic regression model and $\mathbf{Y} | \mathbf{Z}$ is fit based on a negative binomial regression model. The size parameter r is estimated by applying the method of moments to the residuals of the Poisson regression of Y on Z (Barry et al., 2021; Barry et al., 2024). This method (called “precomputed” in Table 1) is fast but less accurate than maximum likelihood estimation, but is sufficient for the spaCRT, which does not require accurate estimation of the size parameter. We use the `uniroot` function in R to solve the equation saddlepoint equation. When the solution is not found or the resulting p -value p_{spaCRT} is not in the range $[0, 1]$, we use the p -value based on the GCM test as a backup (see below). We found the failure to solve the saddlepoint equation quite rare, occurring in at most 1.3% of replications across all simulation settings.
- The **dCRT** (Algorithm 1), with the same fitting procedures as the spaCRT and $M = 10,000$.
- The **GCM test** (Shah and Peters, 2020), which is based on the asymptotically normal test statistic

$$T_n^{\text{GCM}} \equiv \frac{T_n^{\text{dCRT}}(X, Y, Z)}{\widehat{S}_n}, \quad \widehat{S}_n^2 \equiv \frac{1}{n} \sum_{i=1}^n R_{in}^2 - \left(\frac{1}{n} \sum_{i=1}^n R_{in} \right)^2,$$

where $T_n^{\text{dCRT}}(X, Y, Z)$ is defined as in (4) and

$$R_{in} \equiv (X_{in} - \widehat{\mu}_{n,x}(Z_{in}))(Y_{in} - \widehat{\mu}_{n,y}(Z_{in})).$$

We use the same fitting procedures for the GCM test as for spaCRT and dCRT.

- The **negative binomial regression score test** (implemented via the `glm.nb()` function in the MASS package). This function computes the maximum likelihood estimate of the size parameter iteratively, which is slower than the precomputed approach but more accurate. We choose this approach since the score test relies more heavily on the accuracy of the size parameter estimate.

K.2 Additional simulation results in Section 6.1

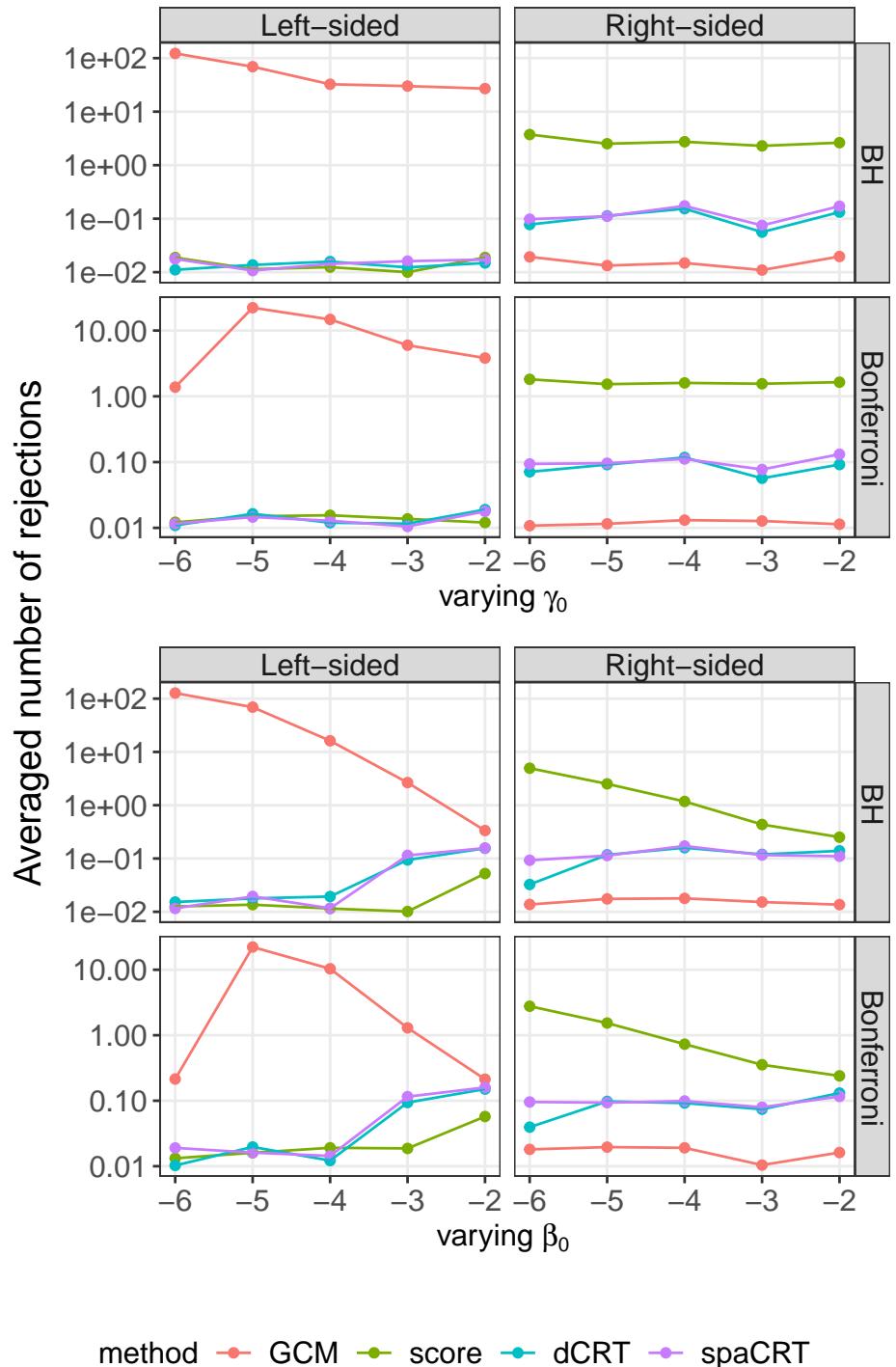


Figure 7: Averaged number of rejections after BH and Bonferroni corrections under the setup $r = 0.05$.

L Additional simulation details in Section 6.2

L.1 Parameters and methods implementation in Section 6.2

Parameters used in Section 6.2 Recall the logistic regression model:

$$\mathcal{L}(\mathbf{Y}|\mathbf{X}) \stackrel{d}{=} \text{Ber}(\text{expit}(\gamma_0 + \mathbf{X}^\top \beta))$$

where γ_0 is an intercept term, β is a vector of coefficient and g is a smooth function. For the concrete choice of β , we consider

$$\beta = (\underbrace{\eta, \dots, \eta}_{0.05*p}, \underbrace{-\eta, \dots, -\eta}_{0.05*p}, \underbrace{0, \dots, 0}_{0.9*p})^\top$$

where $\eta > 0$ is a signal strength. We vary $\eta \in \{0, 0.25, 0.5, 0.75, 1\}$. For γ_0 , we consider $\{-3, -2\}$ for *high* and *low* sparsity settings. For the distribution of \mathbf{X} , we consider the mHMM (Definition 4) where we set the support of the hidden variable \mathbf{U}_j , K , to be 10 and that of \mathbf{X} , M , to be 2. The Markov transition matrix is

$$Q \equiv \begin{pmatrix} \gamma & 1-\gamma & 0 & \dots & 0 & 0 \\ 0 & \gamma & 1-\gamma & \dots & 0 & 0 \\ \vdots & & \ddots & & & \vdots \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \dots & \gamma & 1-\gamma \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix} \in \mathbb{R}^{K \times K}.$$

We set $\gamma = 0.9$ to create non-trivial correlation between \mathbf{X}_j . The intial distribution q is a uniform distribution over the support of \mathbf{U}_1 . Besides the transition matrix, we also vary the emission distribution $\mathbf{X}_j|\mathbf{U}_j$ by considering a beta-piror emission distribution:

$$\mathbb{P}[\mathbf{X}_j = 1|\mathbf{U}_j = k] \stackrel{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, \beta). \quad (\text{beta-emission})$$

Hyperparameter (α, β) controls the shape of the Beta distribution. We consider two choices: $(\alpha, \beta) = (1, 3)$ and $(\alpha, \beta) = (1, 1)$. The first set of parameters will put more mass towards small values close to 0, which induce high sparsity and thus mimic the rare genetic variation setup. The second choice is a uniform distribution over $[0, 1]$ with no skewness. Thus this induce low sparsity in X and thus mimic the common genetic variation scenario. The simulation setup can be summarized in Table 5.

Table 5: Parameters considered in GWAS simulation.

Parameters $(\alpha, \beta, \gamma_0)$	Sparsity level for (X, Y)
$(1, 3, -2)$	(high, low)
$(1, 3, -3)$	(high, high)
$(1, 1, -2)$	(low, low)
$(1, 1, -3)$	(low, high)

We will consider regularization parameters λ used in `glmnet` to be $\lambda = \text{lambda.1se}$ or $\lambda = \text{lambda.min}$.

Methods details in Section 6.2 We devote this section to discussing the computational details of each method and we will discuss a simple yet powerful computational trick to compute the leave-one-out conditional expectations $\widehat{\mathbb{E}}[\mathbf{Y}|\mathbf{X}_{-j}]$ for dCRT, spaCRT and GCM. We first observe the following identity:

$$\mathbb{E}[\mathbf{Y}|\mathbf{X}_{-j}] = \mathbb{E}[\mathbb{E}[\mathbf{Y}|\mathbf{X}]|\mathbf{X}_{-j}]$$

where the outer expectation is taken with respect to the measure $\mathbf{X}_j|\mathbf{X}_{-j}$. If we can estimate the joint distribution of \mathbf{X} from data X and one regression estimate for $\mathbb{E}[\mathbf{Y}|\mathbf{X}]$, computing the conditional expectation $\mathbb{E}[\mathbf{Y}|\mathbf{X}_{-j}]$ for any $j \in [p]$ is straightforward via the integral evaluation with respect to measure $\mathbf{X}_j|\mathbf{X}_{-j}$, without any additional regression fit. In other words, we only need one regression fit for $\mathbb{E}[\mathbf{Y}|\mathbf{X}]$ and one joint distribution fit for \mathbf{X} . In practice, we consider using the following algorithm:

1. **Estimate distribution \mathbf{X} :** this can be done by using `fastPhase` (Scheet and Stephens, 2006);
2. **Compute regression estimate $\widehat{\mathbb{E}}[\mathbf{Y}|\mathbf{X} = x]$:** this can be done by using `glmnet` (Tibshirani, 1996) with the family set to be `binomial`;
3. **Compute $\widehat{\mathbb{E}}[\mathbf{Y}|\mathbf{X}_{-j}]$ for any $j \in [d]$:** this can be done by computing the following integral:

$$\widehat{\mathbb{E}}[\mathbf{Y}|\mathbf{X}_{-j} = x_{-j}] = \sum_{x_j \in \{0,1\}} \widehat{\mathbb{E}}[\mathbf{Y}|\mathbf{X} = x] \widehat{\mathbb{P}}[\mathbf{X}_j = x_j | \mathbf{X}_{-j} = x_{-j}],$$

where $\widehat{\mathbb{P}}[\mathbf{X}_j = x_j | \mathbf{X}_{-j} = \cdot]$ is estimated using the `fastPhase` algorithm.

L.2 Additional simulation results in Section 6.2

We present four simulation results with each corresponding to the choice of `lambda.min` or `lambda.1se` and the sparsity level in X . High sparsity in X with `lambda.1se` has been presented in Figure 3. The other three plots can be found in Figure 8, 9 and 10.

We want to point out that the choice of λ seems to affect both the FDR and power of the methods. In particular, comparing Figure 8 and Figure 9, we can find the FDR can be slightly inflated when `lambda.min` is used for dCRT, GCM and spaCRT methods whereas the FDR is well controlled when `lambda.1se` is used. The inflation of false positive rate can be because of the tower trick used for these methods. The intuition is that when `lambda.1se` is used for dCRT, GCM and spaCRT, the estimated regression coefficients are more sparse and the models obtained from `glmnet` is less variable whereas `lambda.min` will lead to more variable models due to the relatively dense model it will produce. On the power side, we can see that the power of dCRT, GCM and spaCRT is slightly improved when `lambda.min` is used. Interestingly, the FDR of Knockoff procedure seems to be more robust to the choice of λ whereas the power is worse when `lambda.min` is used, which is different from the behavior of dCRT, GCM and spaCRT.

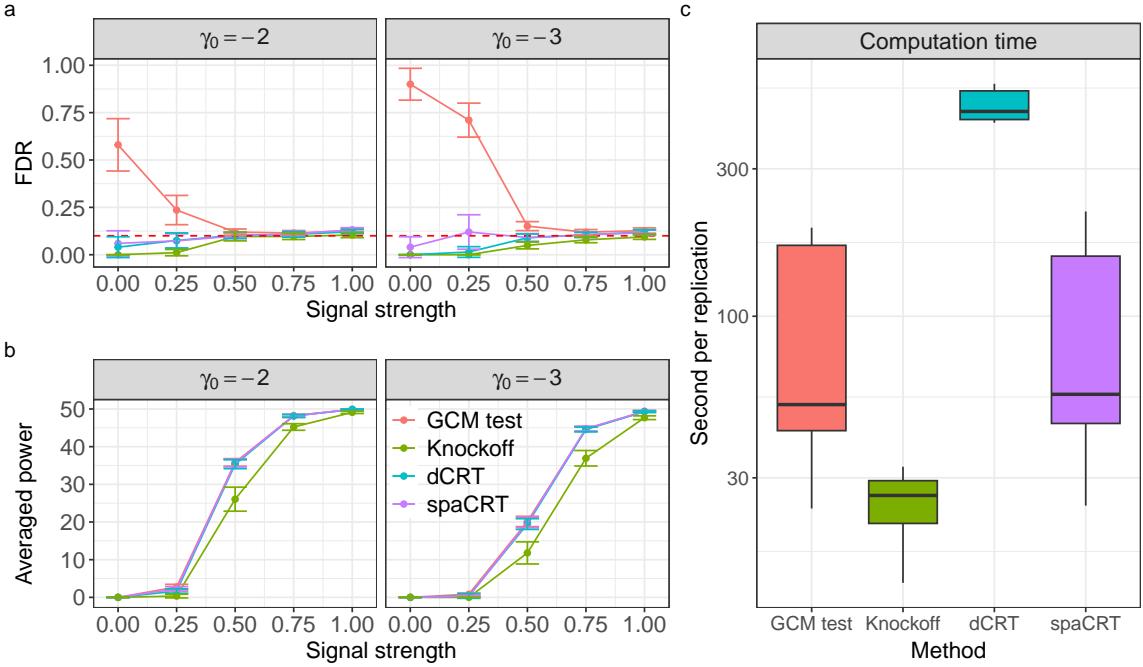


Figure 8: Summary of numerical simulation results for variable selection with low sparsity in data X . All the results are obtained with the regularization parameter $\lambda = \text{lambda.1se}$. (a) FDR for $\gamma_0 = -3$ (high sparsity) and $\gamma_0 = -2$ (low sparsity). (b) Power for the same set of γ_0 . (c) Computation times consumed by different methods.

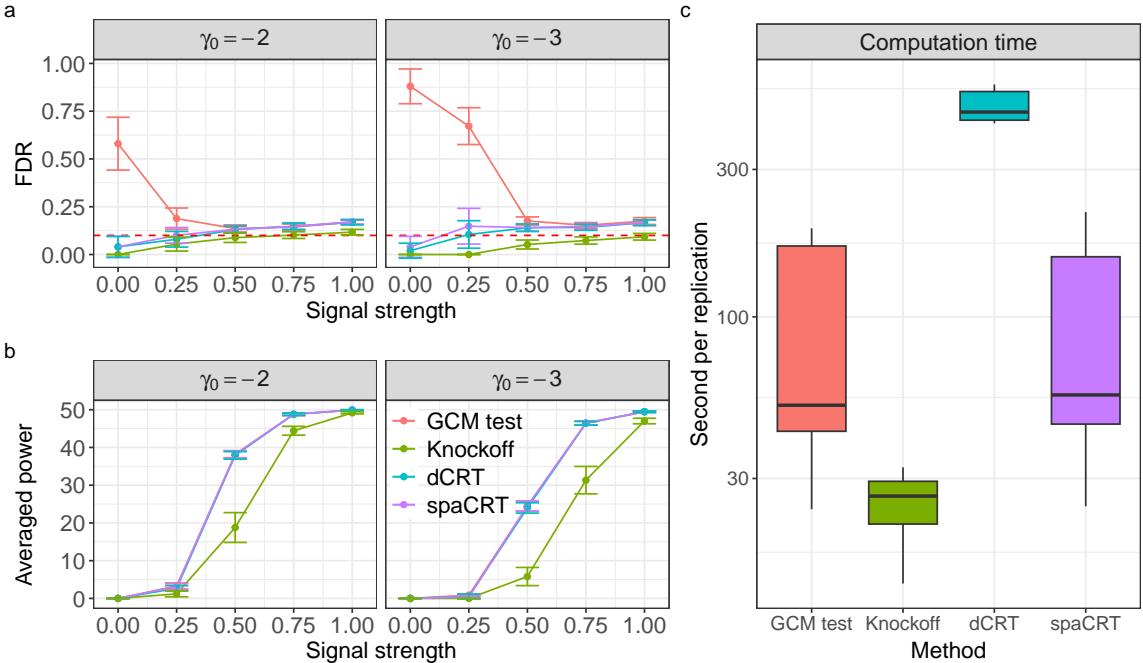


Figure 9: Summary of numerical simulation results for variable selection with low sparsity in data X . All the results are obtained with the regularization parameter $\lambda = \text{lambda.min}$. (a) FDR for $\gamma_0 = -3$ (high sparsity) and $\gamma_0 = -2$ (low sparsity). (b) Power for the same set of γ_0 . (c) Computation times consumed by different methods.

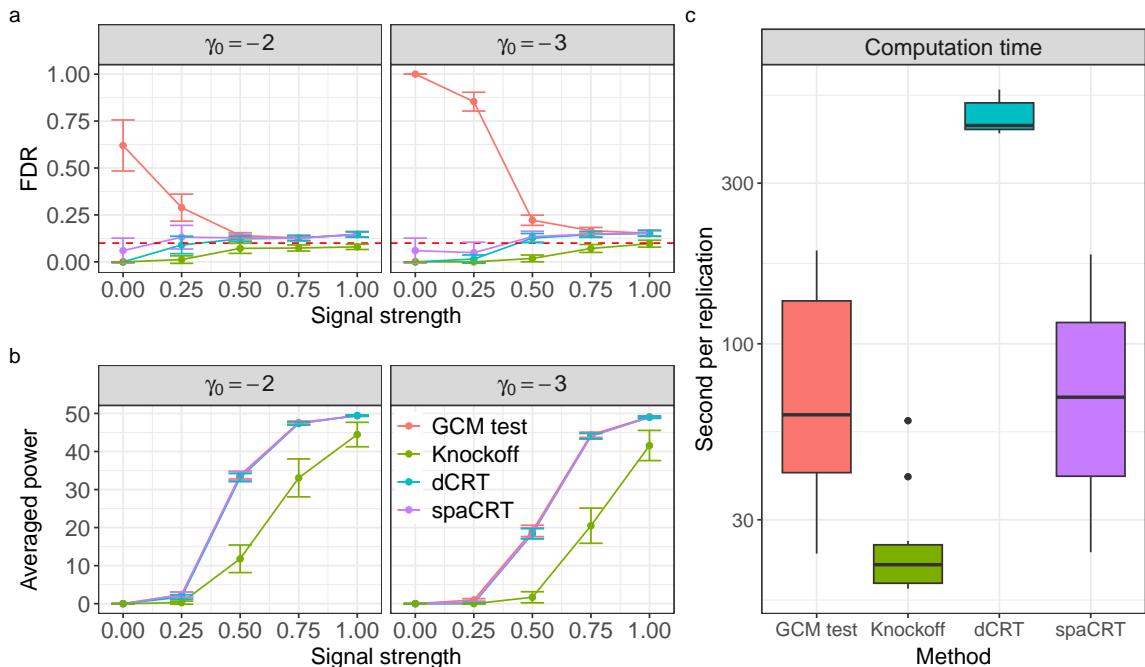


Figure 10: Summary of numerical simulation results for variable selection with high sparsity in data X . All the results are obtained with the regularization parameter $\lambda = \text{lambda}.\min$. (a) FDR for $\gamma_0 = -3$ (high sparsity) and $\gamma_0 = -2$ (low sparsity). (b) Power for the same set of γ_0 . (c) Computation times consumed by different methods.

M Simulation on unbalanced nonparametric classification

N Additional figures and tables for real data analysis

In section N.1, we show a table including the number of rejections when applying Bonferroni or BH method to the pairs involving the non-targeting perturbations (thus under the null). The total number of hypotheses is 153000. In section N.2, we present additional figures for real data analysis including QQ-plots facetting across different effective sample size (Figure 11) and QQ-plots facetting across different dispersion parameters (Figure 12).

N.1 Additional tables for the real data analysis

Table 6: Number of rejections for negative control pairs on the Gasperini data.

Method	Number of rejections			
	Left-sided test		Right-sided test	
	Bonferroni	BH	Bonferroni	BH
GCM test	22	128	0	0
Score test	1	1	15	29
spaCRT	1	1	0	0
dCRT	1	4	0	0

N.2 Additional figures for the real data analysis

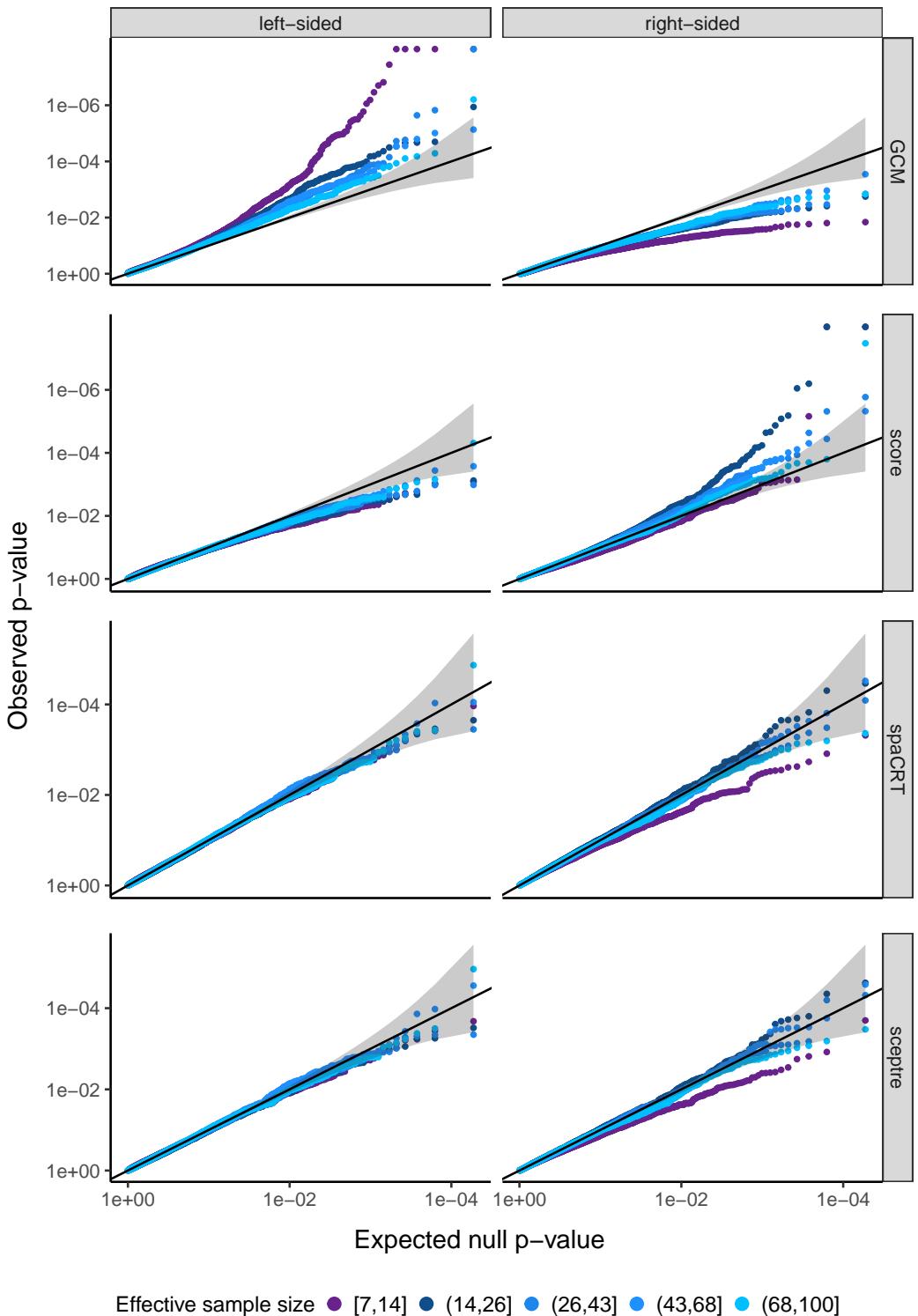


Figure 11: QQ-plots for the p -values of right-sided test from different methods under low effective sample size.

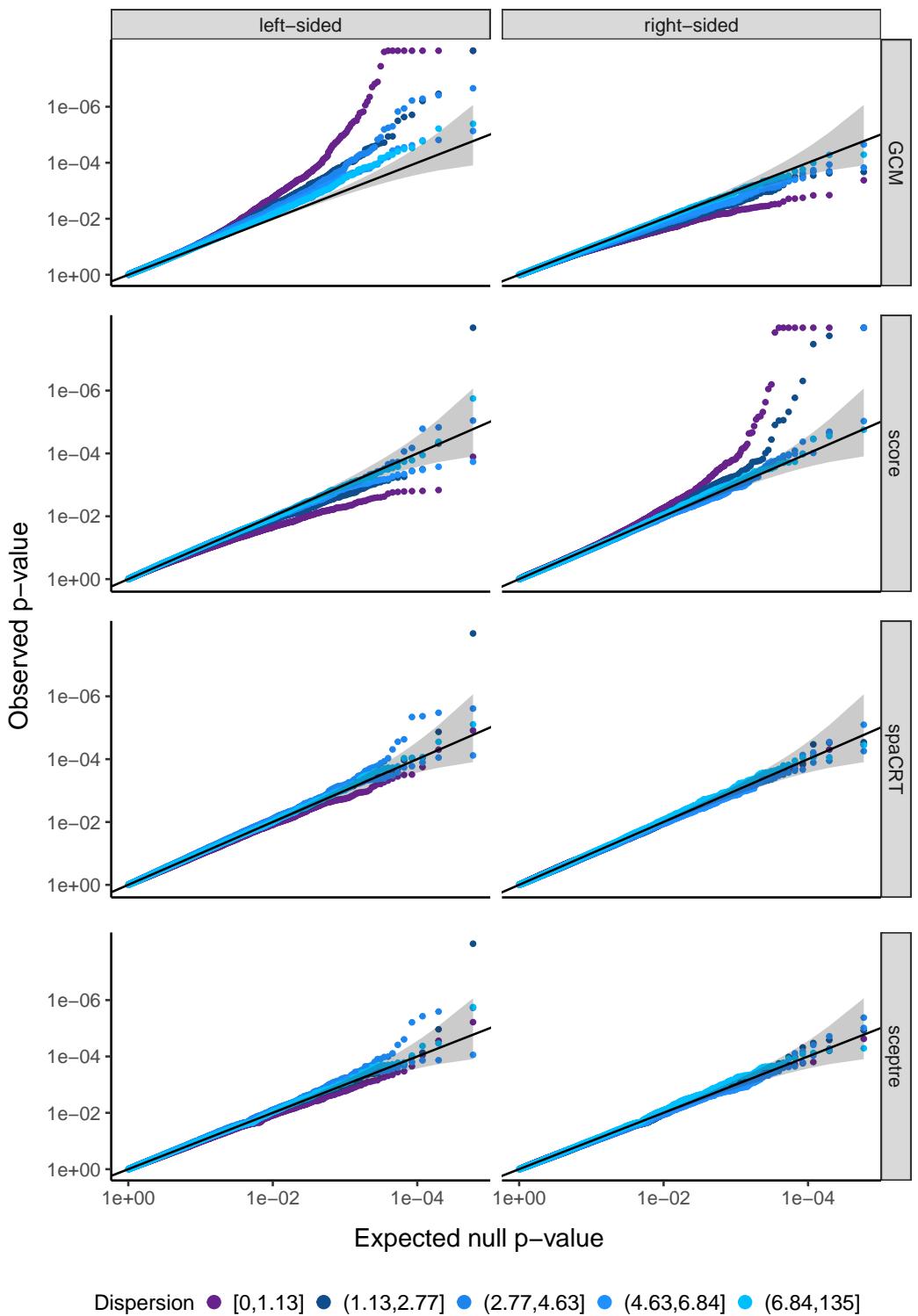


Figure 12: QQ-plots for the p -values of left-sided test from different methods stratified by dispersion parameter.