# Doubly robust and computationally efficient high-dimensional variable selection

Abhinav Chakraborty[*], Jeffrey Zhang[*] and Eugene Katsevich

September 14, 2024

**Abstract**

The variable selection problem is to discover which of a large set of predictors is associated with an outcome of interest, conditionally on the other predictors. This problem has been widely studied, but existing approaches lack either power against complex alternatives, robustness to model misspecification, computational efficiency, or quantification of evidence against individual hypotheses. We present tower PCM (tPCM), a statistically and computationally efficient solution to the variable selection problem that does not suffer from these shortcomings. tPCM adapts the best aspects of two existing procedures that are based on similar functionals: the holdout randomization test (HRT) and the projected covariance measure (PCM). The former is a model-X test that utilizes many resamples and few machine learning fits, while the latter is an asymptotic doubly-robust style test for a single hypothesis that requires no resamples and many machine learning fits. Theoretically, we demonstrate the validity of tPCM, and perhaps surprisingly, the asymptotic equivalence of HRT, PCM, and tPCM. In so doing, we clarify the relationship between two methods from two separate literatures. An extensive simulation study verifies that tPCM can have significant computational savings compared to HRT and PCM, while maintaining nearly identical power.

# 1 Introduction

## 1.1 The variable selection problem

With the advancement of scientific data acquisition technologies and the proliferation of digital platforms collecting user data, it is increasingly common to have databases with large numbers of variables. In this context, a common statistical challenge is variable selection: identifying a subset of predictors that are relevant to a response variable of interest. For example, in genetics, researchers aim to identify genetic variants that influence disease susceptibility, while in finance, analysts seek key indicators that predict market trends. In these problems and many others, only a small subset of the available

[*]Equal contribution.

predictors are expected to have an impact on the response. The task is to identify these important predictors amid a sea of irrelevant ones.

Let us denote the predictor variables $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p) \in \mathbb{R}^p$ and the response variable $\boldsymbol{Y} \in \mathbb{R}$. We are given $m + n$ i.i.d. observations $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} \mathcal{L}(\boldsymbol{X}, \boldsymbol{Y})$, collectively denoted $(X, Y) \equiv \{(X_i, Y_i)\}_{i=1,\ldots,m+n}$. One formulation of the variable selection problem is to test the conditional independence of $\boldsymbol{Y}$ and the predictor $\boldsymbol{X}_j$ given the other predictors $\boldsymbol{X}_{\text{-}j}$, for each $j$ (Candès et al., 2018):

$$H_{0j} : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j} \quad \text{versus} \quad H_{1j} : \boldsymbol{Y} \not\perp\!\!\!\perp \boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j}. \tag{1}$$

Testing these $p$ hypotheses brings both statistical and computational challenges. Statistically, variable selection procedures should control Type-I error under assumptions that are not too stringent, which is difficult due to possibility of confounding by $\boldsymbol{X}_{\text{-}j}$ (Shah and Peters, 2020). Furthermore, variable selection procedures should have power against broad classes of alternatives, given the generality of $H_{1j}$ and the potential complexity of relationships between $\boldsymbol{Y}$ and $\boldsymbol{X}$. Computationally, the challenge is to perform $p$ tests efficiently, especially when $p$ is large.

## 1.2 An overview of existing approaches

There are a number of existing methods to address the variable selection problem, each with its strengths and weaknesses (Table 1). Here, we highlight a selection of existing approaches, while deferring additional discussion of related work to Section 1.4. We evaluate each approach based on four criteria. First, in the variable selection context, we call a method *doubly robust* if it provably controls Type-I error asymptotically while fitting both $\mathcal{L}(\boldsymbol{X})$ and $\mathcal{L}(\boldsymbol{Y} \mid \boldsymbol{X})$ in-sample. This notion is related to, but distinct from the conventional rate double-robustness definition (Smucler, Rotnitzky, and Robins, 2019). Second, a method has power against general alternatives if it has power against alternatives beyond those that can be specified by a single pathwise-differentiable functional $\psi_j(\mathcal{L})$ for each variable $j$, such as the expected conditional covariance

$$\psi_j(\mathcal{L}) \equiv \mathbb{E}_{\mathcal{L}}[\mathrm{Cov}_{\mathcal{L}}[\boldsymbol{X}_j, \boldsymbol{Y} \mid \boldsymbol{X}_{\text{-}j}]]. \tag{2}$$

Third, a method is computationally fast if it does not require running machine learning procedures $p$ times or recomputing the test statistic on $O(p^2)$ resamples. Fourth, a method produces $p$-values for each variable if it quantifies the significance of each variable $\boldsymbol{X}_j$ with a $p$-value, a property that aids with interpretability and is often expected by practitioners.

One class of methods is based on the model-X framework (Candès et al., 2018), which includes model-X knockoffs (Candès et al., 2018) and the holdout randomization test (HRT; Tansey et al., 2022). These methods were developed under the assumption that $\mathcal{L}(\boldsymbol{X})$ is known, which arguably is too strong an assumption except in special cases where this law is under the control of the experimenter (Ham, Imai, and Janson, 2022; Aufiero and Janson, 2022). These methods are often deployed by fitting $\mathcal{L}(\boldsymbol{X})$ in-sample, but general conditions under which the Type-I error is controlled in this context are not known, though some progress in this direction has been made (Fan et al., 2019; Fan, Gao, and Lv, 2023). Both HRT and model-X knockoffs are compatible with a broad

| | Model-X | | Doubly robust | | Best of both |
|---|---|---|---|---|---|
| | Knockoffs | HRT | GCM | PCM | tPCM |
| Doubly robust | ? | ? | ✓ | ✓ | ✓ |
| Power against general alternatives | ✓ | ✓ | | ✓ | ✓ |
| Computationally fast | ✓ | | | | ✓ |
| Produces $p$-values for each variable | | ✓ | ✓ | ✓ | ✓ |

Table 1: Comparison of four existing variable selection methods and the proposed method (tPCM) based on four statistical and computational metrics.

range of test statistics, facilitating power against general alternatives. Model-X knockoffs is computationally fast, but it does not produce $p$-values quantifying the significance of each variable, hampering interpretability in applications. On the other hand, the HRT produces $p$-values, but it is computationally expensive, requiring $O(p^2)$ resamples (each of the $p$ variables requires $O(p)$ resamples to reach significance after multiplicity correction).

Another class of methods is constructed based on product-of-residuals test statistics designed to be doubly robust. This class includes the generalized covariance measure (GCM) test (Shah and Peters, 2020) and the projected covariance measure (PCM; Lundborg et al., 2022). These methods are doubly robust by design. The GCM test is only powerful against alternatives where the expected conditional covariance (2) between $\boldsymbol{Y}$ and $\boldsymbol{X}_j$ given $\boldsymbol{X}_{\text{-}j}$ is non-zero, while the PCM test is an extension of the GCM test that is powerful against a broader class of alternatives. Both of these methods were designed with a single conditional independence testing problem in mind, so their direct application to the variable selection problem is computationally slow due to the requirement of at least one machine learning fit per variable. On the other hand, both methods produce $p$-values for each variable.

## 1.3 Our contributions

As is apparent from Table 1, none of the existing model-X or doubly robust methods for variable selection satisfies all four of the criteria we have outlined. Our primary contribution is to introduce a new method, the tower PCM (tPCM), which integrates ideas from both strands of the literature to overcome their respective limitations. We preview the statistical and computational performance of tPCM in Figure 1, compared to an oracle variant of the GCM test, the HRT, and the PCM test (we exclude model-X knockoffs from this comparison because it is not applicable to family-wise error rate control, which is the setting of our comparison). The tPCM is 1-2 orders of magnitude faster than the PCM test and HRT, while having nearly the same statistical power. On the other hand, the GCM test is much less powerful, since the alternative considered here does not fall under the restricted class of alternatives that the GCM test is powerful against.

The tPCM satisfies each of the criteria considered in Table 1, as we verify using both theory and simulations. In particular, we prove that the tPCM is doubly robust, and we observe Type-I error control in our numerical simulations. By its construction, the tPCM is powered against general alternatives, and this fact is echoed by excellent power in our

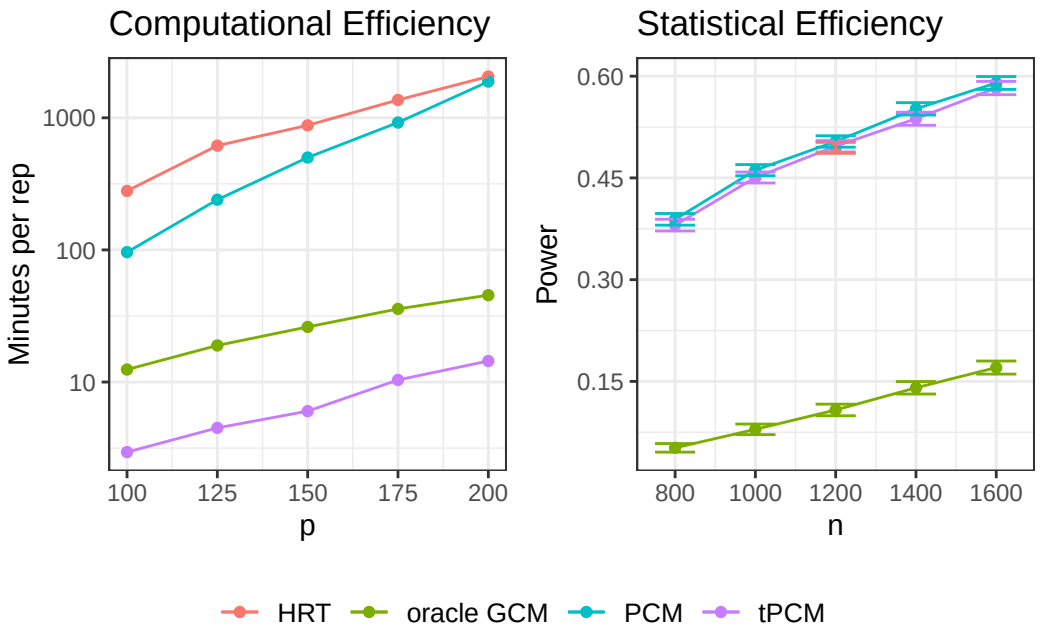


Figure 1: A comparison of the computational and statistical performance of our method, tPCM, with state-of-the-art competitors. The left panel was from a simulation with larger sample size and number of predictors that focused on computational performance, while the right panel was from a simulation that focused on statistical performance.

numerical simulations. We conduct an extensive simulation study to demonstrate the power and computational efficiency of our method, comparing it with existing methods. Additionally, we apply the tPCM to a breast cancer dataset to show its practical applicability. The tPCM only requires two machine learning fits and $O(p)$ resamples, making it computationally fast by our definition. As shown in Figure 1, it is dramatically faster than both the PCM test and the HRT. Finally, the tPCM produces $p$-values for each variable, by its construction. Code to reproduce these simulations and real data analysis is available at `https://github.com/Katsevich-Lab/symcrt2-manuscript/tree/arxiv-v1`.

In addition to our methodological contributions, we provide a novel theoretical insight by showing that the PCM test can be seen as an asymptotic approximation to the HRT. This discovery not only proves the double robustness of the HRT (recall Table 1) but also establishes a bridge between the model-X and doubly robust strands of literature on conditional independence testing and variable selection.

## 1.4 Related work

Here, we expand on different strands of related work.

**Model-X methods.** There have been several works focusing on Type-I error control for model-X methodologies without requiring the model-X assumption. The Type-I error control of model-X knockoffs when $\mathcal{L}(\boldsymbol{X})$ is estimated in-sample has been studied by Fan

et al. (2019) and Fan, Gao, and Lv (2023), as discussed above. This question has also been studied when $\mathcal{L}(\boldsymbol{X})$ is estimated out-of-sample (Barber, Candès, and Samworth, 2020; Fan et al., 2020). A conditional variant of model-X knockoffs that allows $\mathcal{L}(\boldsymbol{X})$ to follow a parametric model with unknown parameters was proposed by Huang and Janson (2020). In addition to model-X knockoffs, Candès et al. (2018) also proposed the conditional randomization test (CRT) for conditional independence testing, of which the HRT is a special case. The Type-I error of the CRT when $\mathcal{L}(\boldsymbol{X})$ is estimated out-of-sample was studied by Berrett et al. (2020). A special case of the CRT called the distilled CRT (dCRT; Liu et al., 2022) was shown to be doubly robust by Niu et al. (2024). Other variants of the CRT have also been proposed for their improved robustness properties (Berrett et al., 2020; Li and Liu, 2022; Barber and Janson, 2022; Zhu and Barber, 2023). Other variants of the CRT have also been proposed for improved computational performance, including the HRT and several others (Tansey et al., 2022; Zhong, Kuffner, and Lahiri, 2021; Li and Candès, 2021; Liu et al., 2022). In the latter category, tests either are not suited for producing fine-grained $p$-values for each variable or require $O(p^2)$ resamples to get them.

**Doubly robust methods.** Another related strand of literature focuses on doubly robust testing and estimation. The GCM test (Shah and Peters, 2020) uses a product of residuals statistic to test conditional independence against alternatives where the expected conditional covariance (2) is nonzero. Minimax estimation of the expected conditional covariance has also been extensively studied; see for example Robins et al. (2008) and Robins et al. (2009). The weighted GCM test (Scheidegger, Hörrmann, and Bühlmann, 2022) extends the GCM test for power against broader classes of alternatives. For sensitivity against even more general departures from the null, estimation and testing of functionals related to

$$\varphi_j(\mathcal{L}) \equiv \mathbb{E}_{\mathcal{L}}[(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}}[\boldsymbol{Y} \mid \boldsymbol{X}_{\text{-}j}])^2] - \mathbb{E}_{\mathcal{L}}[(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}}[\boldsymbol{Y} \mid \boldsymbol{X}])^2] \tag{3}$$

have been considered (Zhang and Janson, 2020; Williamson et al., 2021a; Williamson et al., 2021b; Dai, Shen, and Pan, 2022; Lundborg et al., 2022; Hudson, 2023; Verdinelli and Wasserman, 2024), including the PCM test. The advantage of this functional is that it is equal to zero if and only if conditional independence $H_{0j}$ holds, but its disadvantage is that it is not pathwise differentiable at such points, since they lie on the boundary of the space of values taken on by the functional. Different methods have different approaches to mitigating this issue. However, all of these methods were designed to examine a single variable at a time, so naive application of these approaches to each of the predictor variables is computationally expensive when the number of predictors is large.

**Work at the intersection.** In a previous work (Niu et al., 2024), we established an initial bridge between the model-X and doubly-robust literatures by proving the asymptotic equivalence between two conditional independence tests with power against partially linear alternatives: the dCRT (Liu et al., 2022) and the GCM test (Shah and Peters, 2020). In this work, we strengthen this bridge by proving the asymptotic equivalence between the HRT and the PCM test, which have power against more general classes of alternatives.

# 2 Background: The PCM test and the HRT

In this section, we define the PCM test and the HRT. In preparation for this, we introduce some notation. Let

$$m(\boldsymbol{X}) \equiv \mathbb{E}_{\mathcal{L}}[\boldsymbol{Y} \mid \boldsymbol{X}] \quad \text{and} \quad m_j(\boldsymbol{X}_{\text{-}j}) \equiv \mathbb{E}_{\mathcal{L}}[\boldsymbol{Y} \mid \boldsymbol{X}_{\text{-}j}]. \tag{4}$$

For a fixed function $\widehat{f}(\boldsymbol{X})$, we will denote

$$m_{\widehat{f}}(\boldsymbol{X}_{\text{-}j}) \equiv \mathbb{E}_{\mathcal{L}}[\widehat{f}(\boldsymbol{X}) \mid \boldsymbol{X}_{\text{-}j}]. \tag{5}$$

Many of the quantities we introduce will be indexed by $j$, though at times, we omit this index to lighten notation. We do not assume the model-X setting, so we treat $\mathcal{L}(\boldsymbol{X})$ as unknown. Finally, the set of null laws $\mathcal{L}$ for predictor $j$ is explicitly given by

$$\mathscr{L}^0_{n,j} \equiv \{\mathcal{L} : \mathcal{L}(\boldsymbol{X}_j, \boldsymbol{Y} \mid \boldsymbol{X}_{\text{-}j}) = \mathcal{L}(\boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j}) \times \mathcal{L}(\boldsymbol{Y} \mid \boldsymbol{X}_{\text{-}j})\}. \tag{6}$$

## 2.1 Projected covariance measure

In this section, we describe a "vanilla" version of the PCM methodology proposed in Lundborg et al. (2022), which we shall refer to as vPCM. vPCM is a special case of the slightly more involved PCM, which retains its essential ingredients but omits some steps that do not affect the asymptotic statistical performance. Explicitly, we omit steps 1 (iv) and 2 of Algorithm 1 in Lundborg et al. (2022). The full algorithm is displayed in Algorithm 1, but we describe it in words now. We begin by splitting our data into $D_1 \cup D_2$, with $D_1$ and $D_2$ containing $n$ and $m$ samples, respectively. We estimate $\widehat{m}(\boldsymbol{X}) \equiv \widehat{\mathbb{E}}[\boldsymbol{Y} \mid \boldsymbol{X}]$ on $D_2$, and then we regress it onto $\boldsymbol{X}_{\text{-}j}$ using $D_2$ to obtain $\breve{m}_j(\boldsymbol{X}_{\text{-}j})$. We denote the difference of the two quantities $\widehat{f}_j(\boldsymbol{X}) \equiv \widehat{m}(\boldsymbol{X}) - \breve{m}_j(\boldsymbol{X}_{\text{-}j})$. The quantity $\widehat{f}_j(\boldsymbol{X})$ is then tested for association with $\boldsymbol{Y}$, conditionally on $\boldsymbol{X}_{\text{-}j}$ on $D_1$. To this end, we regress $\boldsymbol{Y}$ on $\boldsymbol{X}_{\text{-}j}$ using $D_1$ to obtain an estimate of $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}_{\text{-}j}]$, which we call $\widetilde{m}_j(\boldsymbol{X}_{\text{-}j})$. We also regress $\widehat{f}_j(\boldsymbol{X})$ on $\boldsymbol{X}_{\text{-}j}$ using $D_1$ to obtain $\widehat{m}_{\widehat{f}_j}(\boldsymbol{X}_{\text{-}j})$. We define the product of residuals stemming from the two regressions as

$$L_{ij} \equiv (Y_i - \widetilde{m}_j(X_{i,\text{-}j}))(\widehat{f}_j(X_i) - \widehat{m}_{\widehat{f}_j}(X_{i,\text{-}j})) \tag{7}$$

and define the vanilla PCM statistic for predictor $j$ as:

$$T_j^{\text{vPCM}} \equiv \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n L_{ij}}{\sqrt{\frac{1}{n} \sum_{i=1}^n L_{ij}^2 - \left(\frac{1}{n} \sum_{i=1}^n L_{ij}\right)^2}} \tag{8}$$

Under the null hypothesis, $T_j^{\text{vPCM}}$ is a sum of random quantities and for sufficiently large $n$ and under appropriate conditions, the Central Limit Theorem (CLT) is expected to apply. Hence, we can compare our statistic to the quantiles of the normal distribution and reject for large values. Our test is defined as

$$\phi_j^{\text{vPCM}}(X, Y) \equiv \mathbb{1}\left(T_j^{\text{vPCM}}(X, Y) > z_{1-\alpha}\right).$$

**Algorithm 1:** Vanilla PCM

**Input:** Data $\{(X_i, Y_i)\}_{i=1,\ldots,m+n}$

Split the data into $D_1 \cup D_2$, with $D_1$ and $D_2$ containing $n$ and $m$ samples, resp.
Estimate $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ on $D_2$, call it $\widehat{m}(\boldsymbol{X})$.
**for** $j \leftarrow 1$ **to** $p$ **do**
    Regress $\widehat{m}(\boldsymbol{X})$ on $\boldsymbol{X}_{\text{-}j}$ using $D_2$ to obtain $\check{m}_j(\boldsymbol{X}_{\text{-}j})$ and define $\widehat{f}_j(\boldsymbol{X}) \equiv \widehat{m}(\boldsymbol{X}) - \check{m}_j(\boldsymbol{X}_{\text{-}j})$.
    Using $D_1$, regress $Y$ on $\boldsymbol{X}_{\text{-}j}$ to obtain an estimate $\widetilde{m}_j(\boldsymbol{X}_{\text{-}j})$ of $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}_{\text{-}j}]$.
    Also on $D_1$, regress $\widehat{f}_j(\boldsymbol{X})$ on $\boldsymbol{X}_{\text{-}j}$ to obtain $\widehat{m}_{\widehat{f}_j}(\boldsymbol{X}_{\text{-}j})$.
    Compute $T_j^{\text{vPCM}}$ based on equations (7) and (8).
    Set $p_j \equiv 1 - \Phi(T_j^{\text{vPCM}})$.
**end**
**return** $\{p_j\}_{j=1,\ldots,p}$.

Aside from the fitting of $\widehat{m}(\boldsymbol{X})$, the steps are repeated for each predictor $j = 1, \ldots, p$.

The primary disadvantage of Algorithm 1 is that it requires $3p+1$ machine learning fits, which we would expect to be computationally difficult when $p$ is large. On the other hand, since Algorithm 1 uses asymptotic approximation, it does not require any resampling.

## 2.2 Holdout Randomization Test

In this section, we describe the holdout randomization test (HRT), displayed as Algorithm 2, which is identical to Algorithm 2 of Tansey et al. (2022) except the estimation of $\mathcal{L}(\boldsymbol{X})$, which the latter authors assumed known. As before, we divide our data into two halves, $D_1$ and $D_2$. On $D_2$, we learn the function $\widehat{m}(\boldsymbol{X}) \equiv \widehat{\mathbb{E}}[\boldsymbol{Y} \mid \boldsymbol{X}]$ and the law $\widehat{\mathcal{L}}(\boldsymbol{X})$. On $D_1$, we compute the mean-squared error (MSE) test statistic

$$T^{\text{HRT}}(X, Y) \equiv \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{m}(X_i))^2. \tag{9}$$

Next, we exploit the fact that under the null hypothesis, the conditional distribution $\mathcal{L}(\boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j}, \boldsymbol{Y})$ is the same as $\mathcal{L}(\boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j})$, for which we have the estimate $\widehat{\mathcal{L}}(\boldsymbol{X}_j|\boldsymbol{X}_{\text{-}j})$. Therefore, we can approximate the distribution of $T^{\text{HRT}}(X, Y)$ conditional on $Y, X_{\text{-}j}, D_2$ by resampling $\widetilde{X}_i \overset{\text{ind}}{\sim} \widehat{\mathcal{L}}(X_{i,j}|X_{i,\text{-}j})$ $B_{\text{HRT}}$ times for each $i = 1, \ldots, n$. In particular, we can approximate the following conditional quantile:

$$C_j(Y, X_{\text{-}j}) \equiv \mathbb{Q}_{1-\alpha} \left[ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{m}(\widetilde{X}_{i,j}, X_{i,\text{-}j}))^2 \mid Y, X_{\text{-}j}, D_2 \right].$$

The HRT for predictor $j$ is then defined as

$$\phi_j^{\text{HRT}}(X, Y) \equiv \mathbb{1}\left(T^{\text{HRT}}(X, Y) \leqslant C_j(Y, X_{\text{-}j})\right).$$

**Algorithm 2:** Holdout Randomization Test

**Input:** Data $\{(X_i, Y_i)\}_{i=1,\dots,m+n}$, $B_{\text{HRT}}$ resamples.

Split the data into $D_1 \cup D_2$, with $D_1$ and $D_2$ containing $n$ and $m$ samples, resp.
Estimate $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ on $D_2$, call it $\widehat{m}(\boldsymbol{X})$.
Estimate $\mathcal{L}(\boldsymbol{X})$ on $D_2$, call it $\widehat{\mathcal{L}}(\boldsymbol{X})$.
Compute test statistic $T^{\text{HRT}}$ as in equation (9).
**for** $j \leftarrow 1$ **to** $p$ **do**
  **for** $b \leftarrow 1$ **to** $B_{\text{HRT}}$ **do**
    Sample $\widetilde{X}_{i,j} \sim \widehat{\mathcal{L}}(X_j \mid X_{i,\text{-}j})$ for all $i \in D_1$.
    Compute $\widetilde{T}_j^b \equiv \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}(\widetilde{X}_{i,j}, X_{i,\text{-}j}))^2$.
  **end**
  Set $p_j \equiv \frac{1}{B_{\text{HRT}}+1}\left(1 + \sum_{b=1}^{B_{\text{HRT}}} \mathbb{1}\left[T^{\text{HRT}} \leqslant \widetilde{T}_j^b\right]\right)$.
**end**
**return** $\{p_j\}_{j=1,\dots,p}$.

The steps are than repeated for each predictor $j = 1, \dots, p$. Algorithm 2 describes how to compute the HRT $p$-values for each variable.

The primary disadvantage of Algorithm 2 is that it requires $p \times B_{\text{HRT}}$ resamples. $B_{\text{HRT}}$ would be required to be large when using the Bonferroni correction to control the family wise error rate, and $p$ is large. On the other hand, an attractive property of Algorithm 2 is that it requires only two machine learning fits.

# 3 Best of both worlds: Tower PCM

In this section, we introduce the tower PCM method (Section 3.1), followed by a discussion of its computational and statistical properties (Sections 3.2 and 3.3, respectively).

## 3.1 The tower PCM algorithm

The computational bottleneck in the application of the PCM test (Algorithm 1) is the repeated application of regressions to obtain $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}_{\text{-}j}]$ for each $j$. Our key observation is that if we compute estimates $\widehat{\mathcal{L}}(\boldsymbol{X})$ and $\widehat{m}(\boldsymbol{X}) \equiv \widehat{\mathbb{E}}[\boldsymbol{Y} \mid \boldsymbol{X}]$ (as in the first two steps of the HRT), then we can construct estimates of $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}_{\text{-}j}]$ for each $j$ without doing any additional regressions. Indeed, note that by the tower property of expectation, we have

$$\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}_{\text{-}j}] \equiv \mathbb{E}_{\mathcal{L}}[m(\boldsymbol{X}) \mid \boldsymbol{X}_{\text{-}j}] \approx \mathbb{E}_{\mathcal{L}}[\widehat{m}(\boldsymbol{X}) \mid \boldsymbol{X}_{\text{-}j}, D_2] \approx \mathbb{E}_{\widehat{\mathcal{L}}}[\widehat{m}(\boldsymbol{X}) \mid \boldsymbol{X}_{\text{-}j}, D_2] \equiv \widehat{m}_j(\boldsymbol{X}_{\text{-}j}).$$

To compute the quantity $\widehat{m}_j$, we can use conditional resampling based on $\widehat{\mathcal{L}}(\boldsymbol{X}_j \mid \boldsymbol{X}_{\text{-}j})$. Unlike the HRT, however, the goal of conditional resampling is to compute expectations rather than tail probabilities, and therefore, much fewer conditional resamples are required. Equipped with $\widehat{m}_j$, we can proceed as in the PCM test by computing products of residuals

$$R_{ij} \equiv (Y_i - \widehat{m}_j(X_{i,\text{-}j}))(\widehat{m}(X_i) - \widehat{m}_j(X_{i,\text{-}j})), \tag{10}$$

and constructing the test statistic

$$T_j^{\text{tPCM}} \equiv \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n R_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^n R_{ij}^2 - \left(\frac{1}{n}\sum_{i=1}^n R_{ij}\right)^2}} \equiv \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n R_{ij}}{\widehat{\sigma}_n}, \tag{11}$$

which we expect is asymptotically normal under the null hypothesis. This yields the test

$$\phi_j^{\text{tPCM}}(X, Y) \equiv \mathbb{1}\left(T_j^{\text{tPCM}}(X, Y) > z_{1-\alpha}\right). \tag{12}$$

These steps lead to Algorithm 3.

**Algorithm 3:** Tower PCM

**Input:** Data $\{(X_i, Y_i)\}_{i=1,\ldots,m+n}$, $B_{\text{tPCM}}$ resamples.

Split the data into $D_1 \cup D_2$, with $D_1$ and $D_2$ containing $n$ and $m$ samples, resp.
Estimate $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ on $D_2$, call it $\widehat{m}(\boldsymbol{X})$.
Estimate $\mathcal{L}(\boldsymbol{X})$ on $D_2$, call it $\widehat{\mathcal{L}}(\boldsymbol{X})$.
**for** $j \leftarrow 1$ **to** $p$ **do**
  **for** $k \leftarrow 1$ **to** $B_{\text{tPCM}}$ **do**
    Sample $\widetilde{X}_{i,j} \sim \widehat{\mathcal{L}}(X_j \mid X_{i,\text{-}j})$ for all $i$.
  **end**
  Compute $\widehat{m}_j(X_{i,\text{-}j}) \equiv \frac{1}{B_{\text{tPCM}}}\sum_{k=1}^{B_{\text{tPCM}}} \widehat{m}(\widetilde{X}_{i,j}, X_{i,\text{-}j})$ for all $i$.
  Define $R_{ij} \equiv (Y_i - \widehat{m}_j(X_{i,\text{-}j}))(\widehat{m}(X_i) - \widehat{m}_j(X_{i,\text{-}j}))$ for $i$ in $D_1$.
  Compute $T_j^{\text{tPCM}} \equiv \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n R_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^n R_{ij}^2 - \left(\frac{1}{n}\sum_{i=1}^n R_{ij}\right)^2}}$.
  Set $p_j \equiv 1 - \Phi(T_j^{\text{tPCM}})$.
**end**
**return** $\{p_j\}_{j=1,\ldots,p}$.

## 3.2 Computational cost comparison

In this subsection, we compare the computational cost of tPCM to that of PCM and HRT. To this end, we consider the following units of computation, which compose the methods considered:

1. `ML(Y|X)`: Training a machine learning model to predict $Y$ from $X$ or $X_{\text{-}j}$
2. `ML(X)`: Training a machine learning model to learn the joint distribution of $X$
3. `ML(X_j|X_-j)`: Training a machine learning model to predict $X_j$ from $X_{\text{-}j}$
4. `predict(X_j|X_-j)`: Sampling or predicting from the conditional distribution of $X_j$ given $X_{\text{-}j}$
5. `predict(Y|X)`: Predicting $Y$ from $X$ using a trained machine learning model

We can consider each method as having a *learning step* (involving some combination of items 1-3) followed by a *prediction step* (involving some combination of items 4-5). Table 2 summarizes the number of each units of computation required by each method, in terms of the number of variables $p$, the number of resamples for tPCM $B_{\text{tPCM}}$, and the number of resamples for HRT $B_{\text{HRT}}$.

| | `ML(Y\|X)` | `ML(X)` | `ML(X_j\|X_-j)` | `predict(X_j\|X_-j)` | `predict(Y\|X)` |
|---|---|---|---|---|---|
| tPCM | 1 | 1 | 0 | $p \times B_{\text{tPCM}}$ | $p \times B_{\text{tPCM}}$ |
| PCM | $p$ | 0 | $p$ | $p$ | $p$ |
| HRT | 1 | 1 | 0 | $p \times B_{\text{HRT}}$ | $p \times B_{\text{HRT}}$ |

Table 2: Computational work required by the methods considered.

Given Table 2, tPCM has a substantial computational advantage over PCM under the following mild conditions:

- $p \gg 1$
  (tPCM has faster machine learning step for $Y|X$)
- $p \times$ (`ML(Y|X)` + `ML(X_j|X_-j)`) $\gg$ `ML(X)`
  (Fitting `ML(X)` for tPCM is negligible compared to repeating ML $p$ times for PCM)
- `ML(X_j|X_-j)` + `ML(Y|X)` $\gg B_{\text{tPCM}} \times$ (`predict(X_j|X_-j)` + `predict(Y|X)`)
  (tPCM's slower prediction step negligible compared to PCM's ML step)

On the other hand, tPCM has a substantial computational advantage over HRT under the following mild conditions:

- $B_{\text{HRT}} \gg B_{\text{tPCM}}$
  (tPCM has faster prediction step)
- $p \times B_{\text{HRT}} \times$ (`predict(X_j|X_-j)` + `predict(Y|X)`) $\geqslant$ `ML(X)` + `ML(Y|X)`
  (HRT's slower prediction step is a significant proportion of its total computation)

In summary, we anticipate that tPCM has a faster machine learning step than PCM and a faster prediction step than HRT. Now that we have established the potential computational advantages of tPCM, we verify its statistical validity.

## 3.3 Type-I error control and equivalence to the oracle test

In this section, we establish the Type-I error control of the tPCM test. To this end, we will show that the tPCM test is asymptotically equivalent to an oracle test. For the remainder of this section, we will focus on the test of $H_{0j}$ for a single predictor $j$, and sometimes omit the index $j$ to lighten the notation. To define the oracle test, we begin by defining the residuals

$$\varepsilon_i \equiv Y_i - m(X_{i,\text{-}j}) \quad \text{and} \quad \xi_i = \widehat{m}(X_{i,j}, X_{i,\text{-}j}) - \mathbb{E}_{\mathcal{L}}[\widehat{m}(X_{i,j}, X_{i,\text{-}j})|X_{i,\text{-}j}, D_2], \tag{13}$$

Note that $\xi_i$ is defined in terms of the estimated $\widehat{m}$ rather than the true $m$. The "oracle" portion consists of access to the true $\mathcal{L}(\boldsymbol{X})$ to compute the conditional expectation term. Letting

$$\sigma_n^2 \equiv \mathrm{Var}_{\mathcal{L}}[\boldsymbol{\varepsilon}\boldsymbol{\xi}|D_2], \tag{14}$$

the oracle test is defined as

$$\phi_j^{\text{oracle}}(X,Y) \equiv \mathbb{1}\left(T_j^{\text{oracle}}(X,Y) > z_{1-\alpha}\right), \quad \text{where} \quad T_j^{\text{oracle}} \equiv \frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^{n}\varepsilon_i\xi_i. \tag{15}$$

Next, we define the asymptotic equivalence of two tests $\phi_n^{(1)}, \phi_n^{(2)} : (X,Y) \mapsto [0,1]$ as the statement

$$\lim_{n\to\infty}\mathbb{P}[\phi_n^{(1)}(X,Y) \neq \phi_n^{(2)}(X,Y)] = 0.$$

The following set of properties will ensure the equivalence of $\phi_j^{\text{tPCM}}$ and $\phi_j^{\text{oracle}}$. The first condition bounds the conditional variance of the error $\varepsilon_i$:

$$\exists c_1 > 0, \quad \mathbb{P}\left[\max_{i\in[n]}\ \mathrm{Var}_{\mathcal{L}}(\varepsilon_i|X_{i,\text{-}j}, D_2) \leqslant c_1\right] \to 1. \tag{16}$$

The next condition is written in terms of the conditional chi-square divergence

$$\chi^2(P,Q \mid \mathcal{F}) \equiv \mathbb{E}_Q\left[\left(\frac{dP}{dQ} - 1\right)^2 \mid \mathcal{F}\right], \tag{17}$$

defined for measures $P$ and $Q$ and a $\sigma$-algebra $\mathcal{F}$. Using the conditional chi-square divergence to measure the error in the conditional distribution $\mathcal{L}_{\boldsymbol{X}_j|\boldsymbol{X}_{\text{-}j}}$, we assume this conditional distribution is consistently estimated in the following sense:

$$\mathbb{P}\left(\max_{i\in[n]}\chi^2\left(\widehat{\mathcal{L}}_{X_{i,j}|X_{i,\text{-}j}}, \mathcal{L}_{X_{i,j}|X_{i,\text{-}j}}|D_2\right) < c_3\right) \to 1, \tag{18}$$

$$E_{\widehat{\mathcal{L}},n}^2 \equiv \frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\chi^2\left(\widehat{\mathcal{L}}_{X_{i,j}|X_{i,\text{-}j}}, \mathcal{L}_{X_{i,j}|X_{i,\text{-}j}}|D_2\right)\mathbb{E}_{\mathcal{L}}[\xi_i^2|X_{i,\text{-}j}, D_2] \xrightarrow{p} 0. \tag{19}$$

Similarly, we assume a consistent estimate of $m(\boldsymbol{X}) = m_j(\boldsymbol{X}_{\text{-}j})$ (this equality holding because we are under the null):

$$(E'_{\widehat{m},n})^2 \equiv \frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{L}}[(\widehat{m}(X_{i,j}, X_{i,\text{-}j}) - m_j(X_{i,\text{-}j}))^2 \mid D_2, X_{i,\text{-}j}]\mathbb{E}[\xi_i^2|X_{i,\text{-}j}, D_2] \xrightarrow{p} 0. \tag{20}$$

Also, we define the MSE for $\widehat{m}$ as follows:

$$E_{\widehat{m},n}^2 = \frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{L}}[(\widehat{m}(X_{i,j}, X_{i,\text{-}j}) - m_j(X_{i,\text{-}j}))^2 \mid D_2, X_{i,\text{-}j}],$$

and assume a doubly robust type assumption which states

$$E_{\widehat{\mathcal{L}},n} \cdot E_{\widehat{m},n} = o_p(n^{-1/2}). \tag{21}$$

Finally, we assume the following Lyapunov-type condition:

$$\frac{1}{\sigma_n^{2+\delta}}\mathbb{E}_{\mathcal{L}}\left[|\boldsymbol{\varepsilon}\boldsymbol{\xi}|^{2+\delta} \mid D_2\right] = o_P(n^{\delta/2}), \tag{22}$$

The following theorem establishes the asymptotic validity of our proposed test under the aforementioned assumptions:

**Theorem 1.** *Given $\mathcal{L} \in \mathscr{L}_n^0$ and estimator $\widehat{m}$ and $\widehat{\mathcal{L}}$ satisfying assumptions* (16)*,* (18)*,* (19)*,* (20)*,* (21) *and* (22)*, we have that $\phi_j^{\text{tPCM}}$ is asymptotically equivalent to $\phi_j^{\text{oracle}}$. Additionally, tPCM is asymptotically level $\alpha$:*

$$\mathbb{E}_{\mathcal{L}}\left[\phi_j^{\text{tPCM}}(X, Y)\right] \to \alpha.$$

Next, we provide a simple example in which the assumptions of Theorem 1 are satisfied.

**Linear Model:** For a fixed $p$ we have $(X_{i,j}, Y_i, X_{i,\text{-}j}) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^p$ for $i = 1, \ldots, 2n$ i.i.d samples arising out of the linear model:

$$\begin{aligned} Y_i &= \beta X_{i,j} + X_{i,\text{-}j}^T\gamma + \epsilon_i \\ X_{i,j} &= X_{i,\text{-}j}^T\eta + \delta_i, X_{i,\text{-}j} \sim P_{\boldsymbol{X}_{\text{-}j}} \end{aligned} \tag{23}$$

where $\epsilon_i, \delta_i \sim N(0,1)$ and $P_{\boldsymbol{X}_{\text{-}j}}$ has bounded support, i.e. $\exists\, c_{\boldsymbol{X}_{\text{-}j}} > 0$ such that $\|\boldsymbol{X}_{\text{-}j}\|_2 \leqslant c_{\boldsymbol{X}_{\text{-}j}}$. We split our data into two halves and estimate all of the unknown parameters using the least squares estimates; this yields estimates $\widehat{m}(\boldsymbol{X}_j, \boldsymbol{X}_{\text{-}j})$ and $\widehat{\mathcal{L}}_{\boldsymbol{X}_j|\boldsymbol{X}_{\text{-}j}}$.

**Lemma 1.** *For the linear model described in* (23)*, under the null (i.e. $\beta = 0$), the assumptions* (16)*,* (18)*,* (19)*,* (20)*,* (21)*, and* (22) *hold true. Therefore, by Theorem 1 we conclude that $\phi_j^{\text{tPCM}}$ is an asymptotically level $\alpha$ test.*

At this point, we have established the computational advantages of tPCM as well as its statistical validity. In the next section, we compare the power of tPCM to those of PCM and HRT.

# 4 Equivalence of tPCM with existing methods

In this section, we will show that tPCM is asymptotically equivalent to the vPCM (Section 4.1) and the HRT (Section 4.2).

## 4.1 Asymptotic equivalence of vPCM and tPCM

To show the equivalence of tPCM and vPCM, we will show that the latter method is equivalent to the oracle test $\phi_j^{\text{oracle}}$ defined in equation (15), which we have shown is equivalent to tPCM (Theorem 1). The conditions under which vPCM is equivalent to

the oracle test echo those under which Lundborg et al. (2022) showed that PCM controls type-I error. Define the in-sample MSE for the two regressions $\widetilde{m}_j$ and $\widehat{m}_{\widehat{f}_j}$ as follows:

$$\mathcal{E}_{\widetilde{m}} = \frac{1}{n}\sum_{i=1}^{n}(\widetilde{m}_j(X_{i,\text{-}j}) - m_j(X_{i,\text{-}j}))^2, \quad \mathcal{E}_{\widehat{m}_{\widehat{f}}} = \frac{1}{n\sigma_n^2}\sum_{i=1}^{n}(\widehat{m}_{\widehat{f}_j}(X_{i,\text{-}j}) - m_{\widehat{f}_j}(X_{i,\text{-}j}))^2.$$

We assume the following consistency conditions for the regression functions $\widetilde{m}_j$ and $\widehat{m}_{\widehat{f}_j}$:

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^{n}(\widetilde{m}_j(X_{i,\text{-}j}) - m_j(X_{i,\text{-}j}))^2\mathbb{E}[\xi_i^2 \mid X_{i,\text{-}j}] \xrightarrow{p} 0. \tag{24}$$

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^{n}(\widehat{m}_{\widehat{f}_j}(X_{i,\text{-}j}) - m_{\widehat{f}_j}(X_{i,\text{-}j}))^2\mathbb{E}[\varepsilon_i^2 \mid X_{i,\text{-}j}] \xrightarrow{p} 0. \tag{25}$$

We also assume a doubly robust condition on the product of MSEs:

$$\mathcal{E}_{\widetilde{m}} \cdot \mathcal{E}_{\widehat{m}_{\widehat{f}}} = o_p(n^{-1}) \tag{26}$$

**Theorem 2.** *Suppose $\mathcal{L}_n \in \mathscr{L}_n^0$ is a sequence of laws satisfying* (22)*,* (24)*,* (25)*, and* (26)*. Then the test $\phi_j^{\text{vPCM}}$ is asymptotically equivalent to the oracle test $\phi_j^{\text{oracle}}$.*

Combining this result with that of Theorem 1, we obtain the following corollary.

**Corollary 1.** *Let $\mathcal{L}_n \in \mathscr{L}_n^0$ be a sequence of laws satisfying* (16)*,* (18)*,* (19)*,* (20)*,* (21)*,* (22)*,* (24)*,* (25)*, and* (26)*. For any sequence $\mathcal{L}_n'$ of alternative distributions contiguous to the sequence $\mathcal{L}_n$, we have that $\phi_n^{\text{vPCM}}$ is equivalent to $\phi_n^{\text{tPCM}}$ against $\mathcal{L}_n'$ i.e.*

$$\lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_n'}\left[\phi_j^{\text{vPCM}}(X,Y) = \phi_j^{\text{tPCM}}(X,Y)\right] = 1.$$

*In particular, these two tests have the same limiting power:*

$$\lim_{n\to\infty} \left\{\mathbb{E}_{\mathcal{L}_n'}\left[\phi_j^{\text{vPCM}}(X,Y)\right] - \mathbb{E}_{\mathcal{L}_n'}\left[\phi_j^{\text{tPCM}}(X,Y)\right]\right\} = 0.$$

Despite equivalence of vPCM and tPCM, we highlight an important distinction between these two methods. tPCM exclusively employs out-of-sample regressions, where the regressions are conducted on a different dataset from which the test statistic is evaluated. In contrast, vPCM utilizes both in-sample and out-of-sample regressions. As was pointed out by Lundborg et al. (2022), relying on in-sample regressions can be advantageous in finite samples. Nevertheless, the effects of this distinction vanish asymptotically.

## 4.2 Asymptotic equivalence of HRT and tPCM

In this section, we establish the asymptotic equivalence between the HRT and tPCM. This is more complicated than establishing the equivalence between vPCM and tPCM, as the HRT is based on resampling rather than an asymptotic approximation, and its test statistic is less closely related to the PCM test statistic. We proceed in three steps. In the first step, we establish a connection between the HRT and tPCM test statistics (Section 4.2.1). In the second step, we show that the difference between the tPCM test statistic and a rescaled variant of the HRT test statistic is asymptotically negligible (Section 4.2.2). In the third step, we show that the HRT cutoff converges to the standard normal cutoff (Section 4.2.3). This leads us to conclude that the HRT and tPCM are asymptotically equivalent (Section 4.2.4).

### 4.2.1 Connecting the HRT and tPCM test statistics

Let us center and scale $T_j^{\text{HRT}}$ so that a limiting distribution can be obtained, and so that the connection with the tPCM test statistic is clearer. Letting

$$\widehat{\xi}_i \equiv \widehat{m}(X_i) - \widehat{m}_j(X_{i,\text{-}j}) \quad \text{and} \quad \widetilde{\xi}_i \equiv \widehat{m}(\widetilde{X}_i) - \widehat{m}_j(X_{i,\text{-}j}), \tag{27}$$

we find that

$$\begin{aligned} T_j^{\text{HRT}} &\equiv \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}(X_i))^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}((Y_i - \widehat{m}_j(X_{i,\text{-}j})) + (\widehat{m}_j(X_{i,\text{-}j}) - \widehat{m}(X_i)))^2 \\ &= -\frac{2\widehat{\sigma}_n}{\sqrt{n}}T_j^{\text{tPCM}} + \frac{1}{n}\sum_{i=1}^{n}(\widehat{m}_j(X_{i,\text{-}j}) - \widehat{m}(X_i))^2 + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}_j(X_{i,\text{-}j}))^2 \\ &\equiv -\frac{2\widehat{\sigma}_n}{\sqrt{n}}T_j^{\text{tPCM}} + \frac{1}{n}\sum_{i=1}^{n}\widehat{\xi}_i^2 + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}_j(X_{i,\text{-}j}))^2. \end{aligned} \tag{28}$$

Separating the portion depending on $X_j$ from the rest, we find that

$$\begin{aligned} T_j^{\text{HRT}} &= -\frac{2\sigma_n}{\sqrt{n}}\left(\frac{\widehat{\sigma}_n}{\sigma_n}T_j^{\text{tPCM}}(X,Y) - \frac{1}{2\sqrt{n}\sigma_n}\sum_{i=1}^{n}(\widehat{\xi}_i^2 - \mathbb{E}[\widetilde{\xi}_i^2 \mid X_{i,\text{-}j}, D_2])\right) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\widetilde{\xi}_i^2 \mid X_{i,\text{-}j}, D_2] + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}_j(X_{i,\text{-}j}))^2 \\ &\equiv -\frac{2\sigma_n}{\sqrt{n}}T_j^{\text{rHRT}} + g(Y, X_{\text{-}j}, D_2). \end{aligned} \tag{29}$$

We have expressed $T_j^{\text{HRT}}$ as a linear function of a rescaled statistic $T_j^{\text{rHRT}}$, which in turn is a linear function of the tPCM statistic $T_j^{\text{tPCM}}$. Let us define the conditional randomization test based on the re-scaled HRT statistic:

$$\phi_j^{\text{rHRT}}(X,Y) \equiv \mathbb{1}\left(T_j^{\text{rHRT}}(X,Y) > C'_n(Y, X_{\text{-}j})\right), \tag{30}$$

where

$$C'_n(Y, X_{\text{-}j}) \equiv \mathbb{Q}_{1-\alpha}\left[T_j^{\text{rHRT}}(\widetilde{X}_j, X_{\text{-}j}, Y) \mid Y, X_{\text{-}j}, D_2\right]. \tag{31}$$

and the resamples $\widetilde{X}_j$ are generated as in step 7 of the HRT (Algorithm 2). The relationship between $T_j^{\text{HRT}}$ and $T_j^{\text{rHRT}}$ (29) implies that the HRT and rHRT are the same test. Indeed, adding and multiplication by factors not involving $X_j$ will be absorbed by the corresponding transformations of the resampling distributions, leaving the test unchanged. We record this fact in the following lemma.

**Lemma 2.** *We have* $\phi_j^{\text{HRT}}(X,Y) = \phi_j^{\text{rHRT}}(X,Y)$.

### 4.2.2 Bounding the difference between rHRT and tPCM test statistics

Recall from equation (29) that

$$T_j^{\text{rHRT}}(X,Y) \equiv \frac{\widehat{\sigma}_n}{\sigma_n} T_j^{\text{tPCM}}(X,Y) - \frac{1}{2\sqrt{n}\sigma_n}\sum_{i=1}^{n}(\widehat{\xi}_i^2 - \mathbb{E}[\widetilde{\xi}_i^2 \mid X_{i,\text{-}j}, D_2]). \tag{32}$$

We will provide conditions under which the multiplicative factor $\frac{\widehat{\sigma}_n}{\sigma_n}$ tends to one, and the additive term $\frac{1}{2\sqrt{n}\sigma_n}\sum_{i=1}^{n}(\widehat{\xi}_i^2 - \mathbb{E}[\widetilde{\xi}_i^2 \mid X_{i,\text{-}j}, D_2])$ tends to zero. This will imply that the test statistics $T_j^{\text{rHRT}}$ and $T_j^{\text{tPCM}}$ are asymptotically equivalent. The multiplicative factor $\frac{\widehat{\sigma}_n}{\sigma_n}$ tends to one under the assumptions of Theorem 1:

**Lemma 3.** *Under the assumptions of Theorem 1, we have that* $\frac{\widehat{\sigma}_n}{\sigma_n} \xrightarrow{p} 1$.

To show that the additive term tends to zero, we require a few more assumptions:

$$\frac{1}{\sigma_n^2}\mathbb{E}\left[\text{Var}\left(\widehat{\boldsymbol{\xi}}^2 \mid \boldsymbol{X}_{\text{-}j}, D_2\right) \mid D_2\right] = o_p(1). \tag{33}$$

$$\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^{n}(\widehat{m}_j(X_{i,\text{-}j}) - \mathbb{E}\left[\widehat{m}(X_{i,j}, X_{i,\text{-}j}) \mid X_{i,\text{-}j}, D_2\right])^2 \xrightarrow{p} 0. \tag{34}$$

$$\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^{n}\left(\text{Var}_{\widehat{\mathcal{L}}}[\xi_i \mid X_{i,\text{-}j}, D_2] - \text{Var}_{\mathcal{L}}[\xi_i \mid X_{i,\text{-}j}, D_2]\right) \xrightarrow{p} 0. \tag{35}$$

Conditions (34) and (35) can be seen as rate assumptions on the convergence of $\widehat{\mathcal{L}}$ to $\mathcal{L}$. The former equation quantifies the rate of convergence of $\mathbb{E}_{\widehat{\mathcal{L}}}\left[\widehat{m}(\boldsymbol{X}_j, \boldsymbol{X}_{\text{-}j}) \mid \boldsymbol{X}_{\text{-}j}, D_2\right]$ to $\mathbb{E}_{\mathcal{L}}\left[\widehat{m}(\boldsymbol{X}_j, \boldsymbol{X}_{\text{-}j}) \mid \boldsymbol{X}_{\text{-}j}, D_2\right]$ and the latter of $\text{Var}_{\widehat{\mathcal{L}}}[\boldsymbol{\xi} \mid \boldsymbol{X}_{\text{-}j}, D_2]$ to $\text{Var}_{\mathcal{L}}[\boldsymbol{\xi} \mid \boldsymbol{X}_{\text{-}j}, D_2]$.

**Lemma 4.** *Under assumptions* (33)*,* (34) *and* (35)*, we have*

$$\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^{n}(\widehat{\xi}_i^2 - \mathbb{E}[\widetilde{\xi}_i^2 \mid X_{i,\text{-}j}, D_2]) \xrightarrow{p} 0. \tag{36}$$

Now that we have connected the rHRT and tPCM test statistics, we proceed to connect their cutoff values.

### 4.2.3 Convergence of the rHRT cutoff

Here, we provide conditions under which the rHRT cutoff $C_n'(Y, X_{\text{-}j})$ converges to the standard normal cutoff $z_{1-\alpha}$. We assume the following consistency conditions:

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^{n}(\widehat{m}_j(X_{i,\text{-}j}) - m_j(X_{i,\text{-}j}))^2\mathbb{E}(\widetilde{\xi}_i^2 \mid X_{i,\text{-}j}, D_2) \xrightarrow{p} 0. \tag{37}$$

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\left(\text{Var}_{\widehat{\mathcal{L}}}[\xi_i \mid X_{i,\text{-}j}, D_2] - \text{Var}_{\mathcal{L}}[\xi_i \mid X_{i,\text{-}j}, D_2]\right)\mathbb{E}(\varepsilon_i^2 \mid X_{i,\text{-}j}) \xrightarrow{p} 0. \tag{38}$$

$$\frac{1}{\sigma_n^2}\mathbb{E}\left[\text{Var}(\widetilde{\boldsymbol{\xi}}^2 \mid \boldsymbol{X}_{\text{-}j}, D_2) \mid D_2\right] \xrightarrow{p} 0. \tag{39}$$

The first condition can be interpreted as a consistency condition for $\widehat{m}_j$, while the second condition pertains to variance consistency. Additionally, we assume the following moment condition, which is similar to condition (22):

$$\frac{1}{\sigma_n^{2+\delta}}\mathbb{E}(|\boldsymbol{\varepsilon}\widetilde{\boldsymbol{\xi}}|^{2+\delta} \mid D_2) = o_p(n^{\delta}). \tag{40}$$

**Lemma 5.** *Under assumptions* (22), (37), (38), (39), *and* (40), *we have*

$$C_n'(Y, X_{\text{-}j}) \xrightarrow{p} z_{1-\alpha}.$$

At this stage, we can put all of the pieces together to show the asymptotic equivalence of the HRT and tPCM.

### 4.2.4 Asymptotic equivalence of HRT and tPCM

The following theorem can be deduced from Lemmas 2, 3, 4, and 5.

**Theorem 3.** *Suppose $\mathcal{L}_n \in \mathscr{L}_n^0$ is a sequence of laws satisfying the assumptions of Theorem 1, as well as conditions* (33), (34), (35), (37), (38), (39), (40). *Then, the HRT test is equivalent to the tPCM test against $\mathcal{L}_n$.*

One consequence of this theorem is the Type-I error control of the HRT beyond the model-X assumption.

**Corollary 2.** *For a sequence of null laws $\mathcal{L}_n \in \mathscr{L}_n^0$ satisfying the assumptions of Theorem 3, the HRT is asymptotically level $\alpha$.*

Another consequence of Theorem 3 is that HRT and tPCM are equivalent under contiguous alternatives, and therefore have equal asymptotic power against contiguous alternatives.

**Corollary 3.** *If $\mathcal{L}_n'$ is a sequence of alternative distributions contiguous to a sequence $\mathcal{L}_n$ in $\mathscr{L}_n^0$ satisfying the assumptions of Theorem 3, then the HRT and tPCM tests are asymptotically equivalent against $\mathcal{L}_n'$. Furthermore, they have equal asymptotic power against $\mathcal{L}_n'$:*

$$\lim_{n\to\infty}\left\{\mathbb{E}_{\mathcal{L}_n'}[\phi_j^{\mathrm{HRT}}(X,Y)] - \mathbb{E}[\phi_j^{\mathrm{tPCM}}(X,Y)]\right\} = 0. \tag{41}$$

Finally, we claim that the assumptions of Theorem 3 are satisfied in the linear model (23).

**Lemma 6.** *For the linear model described in* (23) *with $\beta = 0$, the assumptions of Theorem 3 are satisfied, which implies that $\phi_j^{\mathrm{tPCM}}$ is an asymptotically equivalent to $\phi_j^{\mathrm{HRT}}$.*

By constructing a null distribution through resampling, the HRT accommodates arbitrarily complex machine learning methods for constructing test statistics, whose asymptotic distributions may not be known. However, we find that after appropriate scaling and centering, the resampling-based null distribution essentially replicates the asymptotic normal distribution utilized by the PCM test. Therefore, when testing a single hypothesis in large samples, the additional computational burden of resampling is unnecessary, as the equivalent PCM test can be applied instead. When dealing with a large number of samples and multiple hypotheses, the tPCM test becomes the natural candidate, combining the best aspects of the existing methodologies. For a small number of samples, the HRT remains an attractive option, as it does not rely on asymptotic normality.

# 5 Finite-sample assessment

In this section, we investigate the finite-sample performance of tPCM with a simulation-based assessment of Type-I error, power, and computation time. The Type-I error of choice was the family-wise error rate at level $\alpha = 0.05$. We consider a generalized additive model (GAM) specification for the distribution of $Y \mid X$. The goal of the simulation is to corroborate the findings of the previous sections: (1) tPCM is computationally efficient, (2) tPCM controls the Type-I error, and (3) tPCM is as powerful as HRT and PCM.

## 5.1 Data-generating model

We pick $s$ of the $p$ variables to be nonnull at random. Let $\mathcal{S}$ denote the set of nonnulls. We then define our data-generating model as follows:

$$\mathcal{L}_n(\mathbf{X}) = N(0, \Sigma(\rho)),\ \mathcal{L}_n(\mathbf{Y} \mid \mathbf{X}) = N\left(\sum_{i \in \mathcal{S}, \text{odd}} (X_i - 0.3)^2/\sqrt{2}\theta + \sum_{i \in \mathcal{S}, \text{even}} -\cos(X_i)\theta, 1\right)$$

Here, the covariance matrix $\Sigma(\rho)$ is an AR(1) matrix with parameter $\rho$; that is, $\Sigma(\rho)_{ij} = \rho^{|i-j|}$. Therefore, the entire data-generating process is parameterized by the five parameters $(n, p, s, \rho, \theta)$; see Table 3. We vary each of the five parameters across five values each, setting the remaining to the default values (in bold).

| $n$ | $p$ | $s$ | $\rho$ | $\theta$ |
|---|---|---|---|---|
| 800 | 30 | 4 | 0.2 | 0.15 |
| 1000 | 40 | 8 | 0.35 | 0.2 |
| **1200** | **50** | **12** | **0.5** | **0.25** |
| 1400 | 60 | 16 | 0.65 | 0.3 |
| 1600 | 70 | 20 | 0.8 | 0.35 |

Table 3: The values of the sample size $n$, covariate dimension $p$, sparsity $s$, autocorrelation of covariates $\rho$, and signal strength $\theta$ used for the simulation study. Each of the parameters $n$, $p$, $s$, $\rho$, $\theta$ was varied among the values displayed in the table while keeping the other four at their default values, indicated in bold. For example, $p = 50$, $s = 12$, $\rho = 0.5$, $\theta = 0.25$ were kept fixed while varying $n \in \{800, 1000, 1200, 1400, 1600\}$.

## 5.2 Methodologies compared

We applied the four methods tPCM, HRT, PCM, and oracle GCM in conjunction with a Bonferroni correction at level $\alpha = 0.05$ to control the family-wise error rate. The PCM implementation was more or less true to Algorithm 1 of Lundborg et al. (2022). For all methods, quantities such as $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ and $\mathbb{E}[f_j(\boldsymbol{X}_j) \mid \boldsymbol{X}_{-j}]$ were fit using (sparse) GAMs. tPCM and HRT exploited knowledge of the banded structure and so $\mathcal{L}(\boldsymbol{X})$ was fit using a banded precision matrix estimate. PCM was also endowed with knowledge of the banded covariance structure. This meant that for any step requiring a $\mathbb{E}[f_j(\boldsymbol{X}_j) \mid \boldsymbol{X}_{-j}]$ fit, we

actually only regressed $f_j(\boldsymbol{X}_j)$ on $\boldsymbol{X}_{j-1}$ and $\boldsymbol{X}_{j+1}$, since $\boldsymbol{X}_j$ is independent of all other $\mathbf{X}_k$ given $\boldsymbol{X}_{j-1}$ and $\boldsymbol{X}_{j+1}$. Oracle GCM was given knowledge of the true $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ and $\mathcal{L}(\boldsymbol{X})$ models. We defer the remaining details of the method implementations to Appendix B.

*Remark* 1. Here we justify the omission of three additional methods from our simulation study. First, we omit the holdout grid test (HGT), a faster version of HRT proposed by Tansey et al. (2022). The HGT employs a discrete, finite grid approximation and a caching strategy. Tansey et al. (2022) theoretically demonstrated the validity of the procedure under the model-X assumption. We chose to omit this method because when the joint distribution of $\boldsymbol{X}$ is not known but estimated, the method may no longer be valid and/or depends on the level of discretization chosen. Furthermore, this method trades off computational resources for memory resources, complicating the comparison. Second, we omit a method proposed by Williamson et al. (2021b) for testing whether the functional (3) equals zero because simulations in Lundborg et al. (2022) demonstrated a sizable gap in power when compared with PCM. Finally, we omitted model-X knockoffs since we desired family-wise error rate control, and model-X knockoffs is not designed to produce fine-grained $p$-values necessary to control this error rate.

## 5.3 Simulation results

Results for power and computation time are presented in Figures 2 and 3 respectively. Figure 4 in Appendix C displays the family-wise error of the methods. Below are our observations from these results:

- As we expect, all methods tend in improve in terms of power as $n$ increases, amplitude increases, $p$ decreases, and $\rho$ decreases. For $s$, there no such monotonic relationship.
- All methods control the family-wise error rate, indicating that in this setting, the $\mathcal{L}(\boldsymbol{Y} \mid \boldsymbol{X})$ and $\mathcal{L}(\boldsymbol{X})$ are learned sufficiently well.
- The oracle GCM has significantly lower power than the other methods, as the test statistic it is based on is most powerful against partially linear alternatives, which is not the case in the simulation design. The other methods have roughly equal power.
- Among the three powerful methods, tPCM is by far the fastest, with the gap widening considerably as $p$ grows.

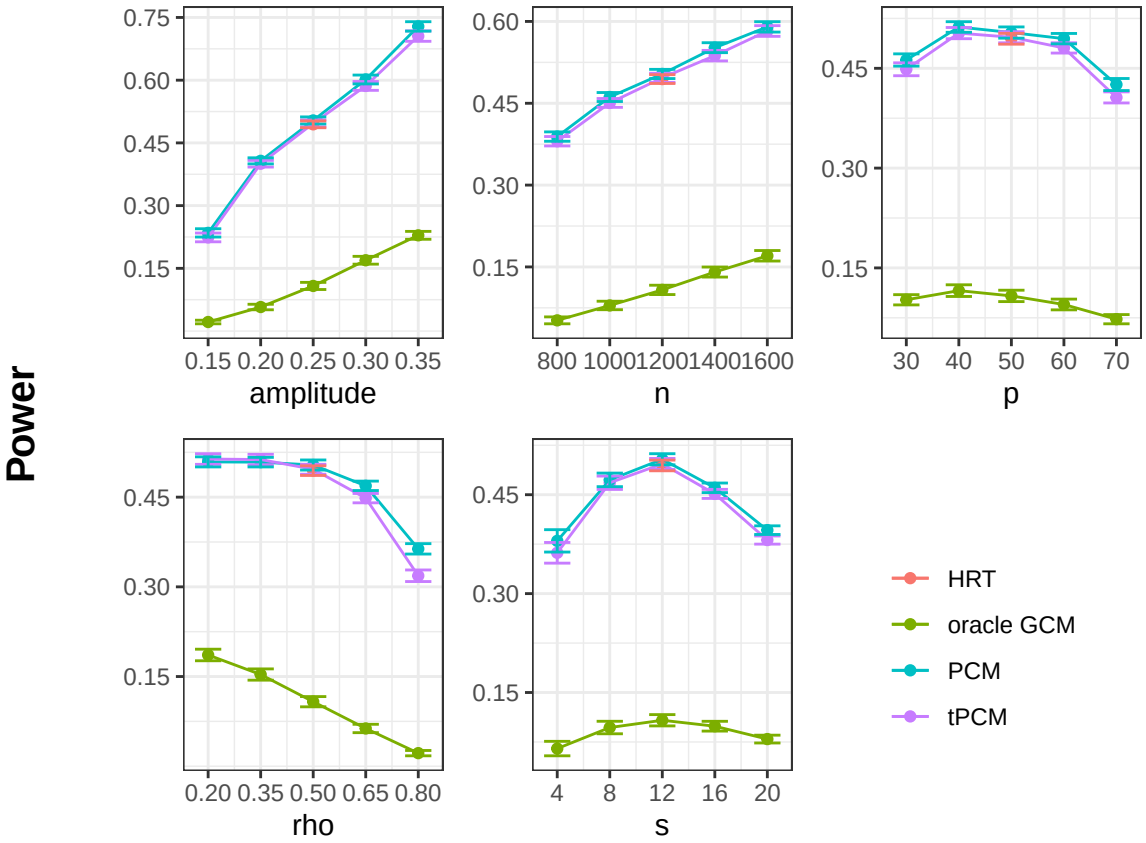


Figure 2: Power: in each plot, we vary one parameter. Each point is the average of 400 Monte Carlo replicates, and the error bars are the average $\pm 2 \times \widehat{\sigma}_p$, where $\widehat{\sigma}_p$ is the Monte Carlo standard deviation divided by $\sqrt{400}$.

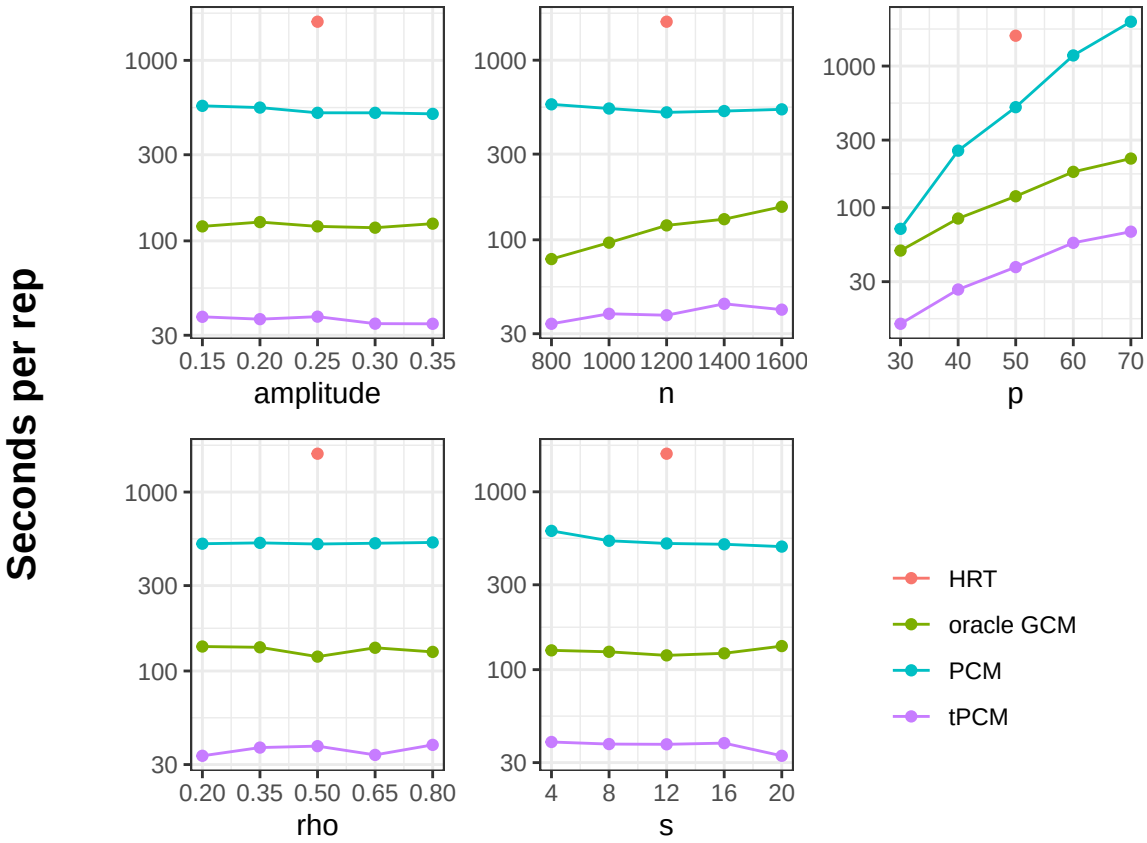


Figure 3: Computation: in each plot, we vary one parameter. Each point is the average of 400 Monte Carlo replicates.

## 5.4 Computational comparison in a larger setting

In the main simulation setting, we chose smaller $n$ and $p$ so that it would be computationally feasible to run 400 Monte Carlo replicates of all methods to assess statistical performance. To further demonstrate the computational advantage of tPCM, we considered a larger setting with the same data-generating model as before, but with different parameters. Specifically, we fixed $n = 2500$, $p = 100$, $\rho = 0.5$, $\theta = 0.25$, $s = 15$, and varied $p \in \{100, 125, 150, 175, 200\}$. We forego any statistical comparison and simply measure the time taken to perform each procedure once for each of the five settings of $p$. HRT, PCM, and tPCM all used a 0.4 training proportion, HRT used $5 \times p/0.05$ resamples, and tPCM and oracle GCM used 25 resamples. These results were already shown in the left panel of Figure 1 in Section 1.3. As expected, the computational gap between tPCM and HRT and PCM widens as $p$ increases, and when $p = 200$, tPCM is more than 130 times faster than HRT and PCM.

# 6 Application to breast cancer dataset

## 6.1 Overview of the data

As a final illustration of our method, we apply tPCM to a breast cancer dataset from Curtis et al. (2012), which has been previously analyzed in the statistical literature by Liu et al. (2022) and Li and Candès (2021). The data consist of $n = 1396$ positive cases of breast cancer categorized by stage (the outcome variable) and $p = 164$ genes, for which the expression level (mRNA) and copy number aberration (CNA) are measured. We seek to discover genes that are associated with stage of breast cancer, conditional on the remaining genes. Statistically, we set the family-wise error rate to be $\alpha = 0.1$. The data is preprocessed using the same steps as in Liu et al. (2022); we refer the reader to Appendix E of Liu et al. (2022) for more details. The stage of cancer outcome is binary, and the gene expression predictors are continuous.

## 6.2 Methods and their implementations

As in the simulation study, we applied four methods to the data, which were HRT, tPCM, PCM, and tower GCM (tGCM). The methods are similar to those from the simulations, and we again fit $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ and $\mathbb{E}[f_j(\boldsymbol{X}_j) \mid \boldsymbol{X}_{-j}]$ using a sparse GAM. One major distinction, however, was that the covariance structure was not banded as in the simulation, so we fit $\mathcal{L}(\boldsymbol{X})$ using the graphical lasso. This also had implications for the $\mathbb{E}[f_j(\boldsymbol{X}_j) \mid \boldsymbol{X}_{-j}]$ fits in PCM. We leave the details of the implementations for each method and their specific hyperparameters to Appendix E, as well as an explanation of the tGCM procedure, which is similar to the oracle GCM procedure from the simulation.

*Remark* 2. Even though the outcome was binary, we chose to fit $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ and $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}_{-j}]$ using a gaussian family GAM. This choice has precedent (Liu et al. (2022) used a non-logistic statistic when implementing the resampling-free version of dCRT), but was also motivated by computational concerns, as a fit for $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ with a GAM with a logit link took 5-6 times as long. Given that PCM took a bit more than 1.5 days to run with the Gaussian link, we estimate that it would have taken over a week to run, which exceeds the maximum run time allowed on our cluster.

## 6.3 Results

We now report the identities of the rejected genes produced by the four methods at level $\alpha = 0.1$ using the Bonferroni correction, as well as the computation time in minutes. These are summarized in Table 4. Notably, tower PCM returns the most rejections with 4, followed by HRT and tower GCM with 3, and none from PCM. In addition, tower PCM performs the best computationally, taking just over 8 minutes due to only estimating two nuisances and using 25 resamples for each predictor. HRT likewise estimated two nuisances, but was slower due to applying a black-box function to 5000 resamples. tGCM was bit a slower than tPCM, since it fit two nuisances 5 times corresponding to each fold on 80% of the data. PCM was by far the slowest since it had to fit regressions $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}_{-j}]$ and $\mathbb{E}[f_j(\boldsymbol{X}_j) \mid \boldsymbol{X}_{-j}]$ $p$ times each, and such steps were more costly than resampling. As for

the discovered genes, tPCM had the most rejections with 4, 3 of which were also rejected by HRT. This gives further evidence to the theoretical claim of asymptotic equivalence between HRT and tPCM. It is possible that if we had used more resamples for HRT, it would have discovered the fourth gene that tPCM did, though of course at a greater computational cost. Given our theoretical result regarding equivalence between tPCM and vPCM, it was surprising to see PCM make no rejections. One potential explanation is that there is a disconnect between the vPCM procedure we analyze and the actual PCM procedure from Lundborg et al. (2022) that we implement, as the latter includes several extra steps. Another possibility is simply that the conditions upon which our theory relies are not satisfied in this instance. Finally, tGCM makes 3 rejections, 2 of which match the rejections made by HRT and tPCM. Recall that tGCM uses cross-fitting and thus does not discard any data when testing, while tPCM and HRT use just 65% of the data for testing. That tPCM makes more rejections than tGCM despite the difference in effective sample size suggests that the functional (3) may be better than the functional (2) for detecting the types of alternatives present in this particular dataset.

| Method | Time (mins) | Discovered genes |
|---|---|---|
| Tower PCM | 8.36 | GPS2, MAP3K13, PPP2CB, RUNX1 |
| PCM | 2264.65 | |
| HRT | 385.81 | GPS2, MAP3K13, RUNX1 |
| Tower GCM | 47.76 | FBXW7, MAP3K13, RUNX1 |

Table 4: A summary of the computation time and number of discoveries for each of the methods with family-wise error rate control at level $\alpha = 0.1$ in the breast cancer dataset.

# 7 Discussion

In this paper, we approached the variable selection problem from the dual perspectives of model-X and doubly robust methodologies, focusing on methods with power against broad classes of alternatives. We proved the equivalence of the model-X HRT and the doubly robust PCM, extending the bridge between model-X and doubly robust methodologies we established in Niu et al. (2024). This equivalence showed the doubly robust nature of the HRT test, which had not been established before. Going beyond drawing connections between these two classes of methodologies, we borrowed ideas from both to propose the significantly faster and equally powerful tPCM test.

The primary limitation of the tPCM test, as well as of the PCM test and HRT, is that all of these methodologies rely on sample splitting. We are not aware of any method that can achieve all four of the properties in Table 1 (doubly robust, powerful against general alternatives, computationally fast, and produces $p$-values for each variable) without sample splitting. Unfortunately, cross-fitting cannot be used in conjunction with sample splitting to boost power in this context, since it leads to dependencies between test statistics from different folds. These dependencies can be captured and accounted for by employing the recently proposed rank-transformed subsampling method (Guo and Shah, 2023), though this method is computationally expensive. Sample splitting reduces the

power of these methods compared to model-X knockoffs, which does not require sample splitting. When $p$-values for each variable are not required, for example when targeting false discovery rate control, model-X knockoffs is more powerful than sample-splitting methods. We leave it to future work to explore whether there is a method that can achieve all four properties in Table 1 without sample splitting.

# 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 in this paper. This work was partially supported by NSF DMS-2113072 and NSF DMS-2310654.

# References

Aufiero, Massimo and Lucas Janson (2022). "Surrogate-based global sensitivity analysis with statistical guarantees via floodgate". In: *arXiv*.

Barber, Rina Foygel, Emmanuel J. Candès, and Richard J. Samworth (2020). "Robust inference with knockoffs". In: *Annals of Statistics* 48.3, pp. 1409–1431. arXiv: 1801.03896.

Barber, Rina Foygel and Lucas Janson (2022). "Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling". In: *Annals of Statistic*.

Berrett, Thomas B, Yi Wang, Rina Foygel Barber, and Richard J Samworth (2020). "The conditional permutation test for independence while controlling for confounders". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 82.1, pp. 175–197.

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.

Curtis, Christina et al. (2012). "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups". In: *Nature* 486.7403, pp. 346–352.

Dai, Ben, Xiaotong Shen, and Wei Pan (2022). "Significance Tests of Feature Relevance for a Black-Box Learner". In: *IEEE Transactions on Neural Networks and Learning Systems*.

Fan, Yingying, Emre Demirkaya, Gaorong Li, and Jinchi Lv (2020). "RANK: Large-Scale Inference With Graphical Nonlinear Knockoffs". In: *Journal of the American Statistical Association* 115.529, pp. 362–379. arXiv: 1709.00092.

Fan, Yingying, Lan Gao, and Jinchi Lv (2023). "ARK: Robust Knockoffs Inference with Coupling". In: *arXiv*. arXiv: 2307.04400.

Fan, Yingying, Jinchi Lv, Mahrad Sharifvaghefi, and Yoshimasa Uematsu (2019). "IPAD: Stable Interpretable Forecasting with Knockoffs Inference". In: *Journal of the American Statistical Association*.

Guo, F Richard and Rajen D Shah (2023). "Rank-transformed subsampling: Inference for multiple data splitting and exchangeable p-values". In: *arXiv*. arXiv: 2301.02739v1.

Ham, Dae Woong, Kosuke Imai, and Lucas Janson (2022). "Using Machine Learning to Test Causal Hypotheses in Conjoint Analysis". In: *arXiv*. arXiv: 2201.08343.

Huang, Dongming and Lucas Janson (2020). "Relaxing the Assumptions of Knockoffs by Conditioning". In: *Annals of Statistics, to appear*. arXiv: 1903.02806.

Hudson, Aaron (2023). "Nonparametric inference on non-negative dissimilarity measures at the boundary of the parameter space". In: arXiv: 2306.07492.

Li, Shuangning and Emmanuel J. Candès (2021). "Deploying the Conditional Randomization Test in High Multiplicity Problems". In: *arXiv*. arXiv: 2110.02422.

Li, Shuangning and Molei Liu (2022). "Maxway CRT: Improving the Robustness of Model-X Inference". In: *arXiv*. arXiv: 2203.06496.

Liu, Molei, Eugene Katsevich, Aaditya Ramdas, and Lucas Janson (2022). "Fast and Powerful Conditional Randomization Testing via Distillation". In: *Biometrika* 109.2, pp. 277–293.

Lundborg, Anton Rask, Ilmun Kim, Rajen D. Shah, and Richard J. Samworth (2022). "The Projected Covariance Measure for assumption-lean variable significance testing". In: *arXiv*. arXiv: 2211.02039.

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*.

Polyanskiy, Yury and Yihong Wu (2023). *Information Theory From Coding to Learning*. First. Cambridge University Press.

Robins, James, Lingling Li, Eric Tchetgen, and Aad van der Vaart (2008). "Higher order influence functions and minimax estimation of nonlinear functionals". In: *Probability and Statistics: Essays in Honor of David A. Freedman*. Beachwood, Ohio, USA: Institute of Mathematical Statistics, pp. 335–421.

Robins, James, Eric Tchetgen Tchetgen, Lingling Li, and Aad van der Vaart (2009). "Semiparametric minimax rates". In: *Electronic Journal of Statistics* 3.none.

Scheidegger, Cyrill, Julia Hörrmann, and Peter Bühlmann (2022). "The Weighted Generalised Covariance Measure". In: *Journal of Machine Learning Research* 23, pp. 1–68. arXiv: 2111.04361.

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. arXiv: 1804.07203.

Smucler, Ezequiel, Andrea Rotnitzky, and James M. Robins (2019). "A unifying approach for doubly-robust l1 regularized estimation of causal contrasts". In: *arXiv*. arXiv: 1904.03737.

Tansey, Wesley, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M. Blei (2022). "The Holdout Randomization Test for Feature Selection in Black Box Models". In: *Journal of Computational and Graphical Statistics* 31.1, pp. 151–162. arXiv: 1811.00645.

Verdinelli, Isabella and Larry Wasserman (2024). "Decorrelated Variable Importance". In: *Journal of Machine Learning Research* 25, pp. 1–27.

Williamson, Brian D, Peter B Gilbert, Marco Carone, and Noah Simon (2021a). "Nonparametric variable importance assessment using machine learning techniques". In: *Biometrics* March 2019, pp. 9–22.

Williamson, Brian D et al. (2021b). "A General Framework for Inference on Algorithm-Agnostic Variable Importance A General Framework for Inference on Algorithm-Agnostic Variable Importance". In: *Journal of the American Statistical Association*.

Zhang, Lu and Lucas Janson (2020). "Floodgate : inference for model-free variable importance". In: *arXiv*, pp. 1–67.

Zhong, Yanjie, Todd Kuffner, and Soumendra Lahiri (2021). "Conditional Randomization Rank Test". In: *arXiv*. arXiv: 2112.00258.

Zhu, Wanrong and Rina Foygel Barber (2023). "Approximate co-sufficient sampling with regularization". In: *arXiv* 1. arXiv: 2309.08063.

# A Proofs

Since all of our theoretical results focus on the hypothesis level, where the $j$th hypothesis to be tested is defined in (1), and since $j$ is fixed for the given hypothesis test, we will simplify our notation for clarity. We denote $\boldsymbol{X} = \boldsymbol{X}_j$ and $\boldsymbol{Z} = \boldsymbol{X}_{\text{-}j}$, and in this notation, we are interested in testing the hypothesis:

$$H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \tag{42}$$

In addition, we drop the $j$ subscripts from all quantities. For functions, instead of $m_j(\boldsymbol{X}_{\text{-}j})$, we use $m(\boldsymbol{Z})$ to denote $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{Z}]$, instead of $\widehat{m}_j(\boldsymbol{X}_{\text{-}j})$, we use $\widehat{m}(\boldsymbol{Z})$, and instead of $\widehat{f}_j(\boldsymbol{X}_{\text{-}j})$, we use $\widehat{f}(\boldsymbol{Z})$. We replace $L_{ij}$ and $R_{ij}$ with $L_i$ and $R_i$. Moreover, instead of indexing tests and test statistics by $j$, we index by $n$. We will be using this notation in all of the subsequent sections.

We also define concretely here certain notions of conditional convergence. The first definition is about conditional convergence in distribution.

**Definition 1.** For each $n$, let $W_n$ be a random variable and let $\mathcal{F}_n$ be a $\sigma$-algebra. Then, we say $W_n$ converges in distribution to a random variable $W$ conditionally on $\mathcal{F}_n$ if $\mathbb{P}\left[W_n \leqslant t \mid \mathcal{F}_n\right] \xrightarrow{p} \mathbb{P}[W \leqslant t]$ for each $t \in \mathbb{R}$ at which $t \mapsto \mathbb{P}[W \leqslant t]$ is continuous. We denote this relation via $W_n \mid \mathcal{F}_n \xrightarrow{d,p} W$.

The next definition is about conditional distribution in probability.

**Definition 2.** For each $n$, let $W_n$ be a random variable and let $\mathcal{F}_n$ be a $\sigma$-algebra. Then, we say $W_n$ converges in probability to a constant $c$ conditionally on $\mathcal{F}_n$ if $W_n$ converges in distribution to the delta mass at $c$ conditionally on $\mathcal{F}_n$ (recall Definition 1). We denote this convergence by $W_n \mid \mathcal{F}_n \xrightarrow{p,p} c$. In symbols,

$$W_n \mid \mathcal{F}_n \xrightarrow{p,p} c \text{ if } \quad W_n \mid \mathcal{F}_n \xrightarrow{d,p} \delta_c.$$

## A.1 Proof of results in Section 3

### A.1.1 Auxiliary Lemmas

**Lemma 7** (Lemma S8 from (Lundborg et al., 2022))**.** *Let $(X_{n,i})_{n\in\mathbb{N},i\in[n]}$ be a triangular array of real-valued random variables and let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a filtration on $\mathcal{F}$. Assume that*

*(i)* $X_{n,1}, \ldots, X_{n,n}$ *are conditionally independent given* $\mathcal{F}_n$*, for each* $n \in \mathbb{N}$*;*

*(ii)* $\mathbb{E}_P\left(X_{n,i} \mid \mathcal{F}_n\right) = 0$ *for all* $n \in \mathbb{N}, i \in [n]$*;*

*(iii)* $\left|n^{-1}\sum_{i=1}^n \mathbb{E}_P\left(X_{n,i}^2 \mid \mathcal{F}_n\right) - 1\right| = o_{\mathcal{P}}(1)$*;*

*(iv) there exists* $\delta > 0$ *such that*

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_P\left(\left|X_{n,i}\right|^{2+\delta} \mid \mathcal{F}_n\right) = o_{\mathcal{P}}\left(n^{\delta/2}\right)$$

*Then* $S_n \equiv n^{-1/2}\sum_{m=1}^n X_{n,m}$ *converges uniformly in distribution to* $N(0,1)$*, i.e.*

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}}\sup_{x\in\mathbb{R}}\left|\mathbb{P}_P\left(S_n \leqslant x\right) - \Phi(x)\right| = 0$$

**Lemma 8** ( Lemma S9 from (Lundborg et al., 2022))**.** *Let* $\left(X_{n,i}\right)_{n\in\mathbb{N}, i\in[n]}$ *be a triangular array of real-valued random variables and let* $\left(\mathcal{F}_n\right)_{n\in\mathbb{N}}$ *be a filtration on* $\mathcal{F}$*. Assume that*

*(i)* $X_{n,1}, \ldots, X_{n,n}$ *are conditionally independent given* $\mathcal{F}_n$ *for all* $n \in \mathbb{N}$*;*

*(ii) there exists* $\delta \in (0,1]$ *such that*

$$\sum_{i=1}^n \mathbb{E}_P\left(\left|X_{n,i}\right|^{1+\delta} \mid \mathcal{F}_n\right) = o_{\mathcal{P}}\left(n^{1+\delta}\right).$$

*Then* $S_n \equiv n^{-1}\sum_{i=1}^n X_{n,i}$ *and* $\mu_{P,n} \equiv n^{-1}\sum_{i=1}^n \mathbb{E}_P\left(X_{n,i} \mid \mathcal{F}_n\right)$ *satisfy* $|S_n - \mu_{P,n}| = o_{\mathcal{P}}(1)$*; i.e., for any* $\epsilon > 0$

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}}\mathbb{P}_P\left(|S_n - \mu_{P,n}| > \epsilon\right) = 0.$$

**Lemma 9** (Lemma 2 of Niu et al. (2024))**.** *Let* $W_n$ *be a sequence of nonnegative random variables and let* $\mathcal{F}_n$ *be a sequence of* $\sigma$*-algebras. If* $\mathbb{E}\left[W_n \mid \mathcal{F}_n\right] \xrightarrow{p} 0$*, then* $W_n \xrightarrow{p} 0$*.*

**Lemma 10** (Asymptotic equivalence of tests)**.** *Consider two hypothesis tests based on the same test statistic* $T_n(X,Y,Z)$ *but different critical values:*

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

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

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

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

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

*then the two tests are asymptotically equivalent:*

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

**Lemma 11.** *We have that*

$$(\widehat{m}(Z_i) - \mathbb{E}_{\mathcal{L}}(\widehat{m}(X_i, Z_i)|Z_i, D_2))^2 \leqslant \chi^2\left(\widehat{\mathcal{L}}_{X_i|Z_i}, \mathcal{L}_{X_i|Z_i}|D_2\right)\mathbb{E}_{\mathcal{L}}[\xi_i^2|Z_i, D_2]$$

*we can show that this implies*

$$(\widehat{m}(Z_i) - m(Z_i))^2 \leqslant 2\left(1 + \chi^2\left(\widehat{\mathcal{L}}_{X_i|Z_i}, \mathcal{L}_{X_i|Z_i}|D_2\right)\right)\mathbb{E}_{\mathcal{L}}[(\widehat{m}(X_i, Z_i) - m(Z_i))^2|Z_i, D_2]$$

*Proof of Lemma 11.* Using the variational representation of chi-squared divergence (see for example equation (7.91) in Polyanskiy and Wu (2023))

$$\chi^2(P, Q) = \sup_g \frac{(\mathbb{E}_P(g) - \mathbb{E}_Q(g))^2}{\mathrm{Var}_Q(g)}. \tag{44}$$

For our purposes we will condition throughout on $D_2$. Fix an $i \in [n]$ and additionally condition on $Z_i$, set $P = \widehat{\mathcal{L}}_{X_i|Z_i}$ and $Q = \mathcal{L}_{X_i|Z_i}$. Next we look at a particular $g \equiv \widehat{m}(X_i, Z_i)$, which implies $\mathbb{E}_Q(g) = \mathbb{E}_{\mathcal{L}_{X_i|Z_i}}[\widehat{m}(X_i, Z_i) \mid D_2] = \mathbb{E}_{\mathcal{L}}[\widehat{m}(X_i, Z_i) \mid Z_i, D_2]$ similarly $\mathbb{E}_P(g) = \widehat{m}(Z_i)$. Observe that $\mathrm{Var}_Q(g) = \mathrm{Var}_{\mathcal{L}_{X_i|Z_i}}(\widehat{m}(X_i, Z_i) \mid D_2) = \mathbb{E}_{\mathcal{L}}(\xi_i^2 \mid Z_i, D_2)$. We denote the conditional chi-squared divergence by $\chi^2\left(\widehat{\mathcal{L}}_{X_i|Z_i}, \mathcal{L}_{X_i|Z_i}|D_2\right)$ which then implies by (44) that

$$(\widehat{m}(Z_i) - \mathbb{E}_{\mathcal{L}}(\widehat{m}(X_i, Z_i)|Z_i, D_2))^2 \leqslant \chi^2\left(\widehat{\mathcal{L}}_{X_i|Z_i}, \mathcal{L}_{X_i|Z_i}|D_2\right)\mathbb{E}_{\mathcal{L}}(\xi_i^2|Z_i, D_2),$$

which verifies the first claim.

We can bound $(\widehat{m}(Z_i) - m(Z_i))^2$ as follows:

$$(\widehat{m}(Z_i) - m(Z_i))^2 \leqslant 2(\widehat{m}(Z_i) - \mathbb{E}_{\mathcal{L}}(\widehat{m}(X_i, Z_i)|Z_i, D_2))^2 + 2(\mathbb{E}_{\mathcal{L}}(\widehat{m}(X_i, Z_i)|Z_i, D_2) - m(Z_i))^2 \tag{45}$$

We have already upper bounded the first term.

Observe that using the fact that $(\mathbb{E}\boldsymbol{X})^2 \leqslant \mathbb{E}\boldsymbol{X}^2$ we have that

$$(\mathbb{E}_{\mathcal{L}}(\widehat{m}(X_i, Z_i)|Z_i, D_2) - m(Z_i))^2 \leqslant \mathbb{E}_{\mathcal{L}}[(\widehat{m}(X_i, Z_i) - m(Z_i))^2|Z_i, D_2]. \tag{46}$$

Also using bias variance decomposition inequality we have

$$\mathbb{E}_{\mathcal{L}}(\xi_i^2 \mid Z_i, D_2) = \mathrm{Var}_{\mathcal{L}}(\widehat{m}(X_i, Z_i) \mid Z_i, D_2) \leqslant \mathbb{E}_{\mathcal{L}}[(\widehat{m}(X_i, Z_i) - m(Z_i))^2|Z_i, D_2]. \tag{47}$$

Combining (46) and (47) with (45) the result follows. □

### A.1.2 Proof of main results

The proof of the next result borrows some crucial ideas from Lundborg et al. (2022) and builds on them.

*Proof of Theorem 1.* $T_n^{\mathrm{tPCM}}$ can be written as $T_N/T_D$ where $T_N = \frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^n R_i$ and $T_D = \widehat{\sigma}_n/\sigma_n$ where $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n R_i^2 - \left(\frac{1}{n}\sum_{i=1}^n R_i\right)^2$. We would show that $T_N \xrightarrow{d} N(0, 1)$ and $T_D \xrightarrow{p} 1$. The first of these results is stated as Lemma 12 below. The second is stated as Lemma 3 in Section 4.2 and is proved below. □

**Lemma 12.** *Under the assumptions of Theorem 1, we have that* $T_N \xrightarrow{d} N(0,1)$.

*Proof.* First we analyze $T_N$ for that we decompose $T_N$ into four terms as follows:

$$T_N = \underbrace{\frac{1}{\sqrt{n}\sigma_n}\sum \varepsilon_i \xi_i}_{G_n} - \underbrace{\frac{1}{\sqrt{n}\sigma_n}\sum \varepsilon_i(\widehat{m}(Z_i) - \mathbb{E}_{\mathcal{L}}[\widehat{m}(X_i, Z_i)|Z_i, D_2])}_{A_n} - \underbrace{\frac{1}{\sqrt{n}\sigma_n}\sum \xi_i(\widehat{m}(Z_i) - m(Z_i))}_{B_n}$$
$$+ \underbrace{\frac{1}{\sqrt{n}\sigma_n}\sum (\widehat{m}(Z_i) - m(Z_i))(\widehat{m}(Z_i) - \mathbb{E}_{\mathcal{L}}[\widehat{m}(X_i, Z_i)|Z_i, D_2])}_{C_n}$$

**Term** $G_n$ We use Lemma 7, $\varepsilon_i \xi_i$ are conditionally independent given $\mathcal{F}_n \equiv \sigma(D_2)$. Also note that under the null conditional on $\mathcal{F}_n$, $\varepsilon_i \xi_i / \sigma_n$ are identically distributed random variables with mean zero and unit variance. Hence if we assume (assumption (22)) that

$$\frac{1}{\sigma_n^{2+\delta}} \mathbb{E}_{\mathcal{L}}\left[ |\boldsymbol{\varepsilon}\boldsymbol{\xi}|^{2+\delta} \mid D_2 \right] = o_P(n^{\delta/2})$$

we have that $G_n \xrightarrow{d} N(0,1)$.

**Term** $A_n$ By Lemma 9 it is enough to show $\mathbb{E}[A_n^2 \mid Z, D_2] \xrightarrow{p} 0$. Using the fact that conditionally on $Z, D_2$ the summands of $A_n$ are mean zero and independent we have that it is sufficient to show

$$\mathbb{E}_{\mathcal{L}}[A_n^2 | Z, D_2] \xrightarrow{p} 0 \iff \frac{1}{n\sigma_n^2}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{L}}[\varepsilon_i^2 | Z_i, D_2](\widehat{m}(Z_i) - \mathbb{E}_{\mathcal{L}}[\widehat{m}(X_i, Z_i)|Z_i, D_2])^2 \xrightarrow{p} 0,$$

Using Lemma 11 we have that the above display is implied by

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^{n} \chi^2\left(\widehat{\mathcal{L}}_{X_i|Z_i}, \mathcal{L}_{X_i|Z_i}|D_2\right) \mathbb{E}_{\mathcal{L}}[\xi_i^2|Z_i, D_2]\mathbb{E}_{\mathcal{L}}[\varepsilon_i^2|Z_i] \xrightarrow{p} 0.$$

Next we use assumption (16) to conclude that it is sufficient to have

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^{n} \chi^2\left(\widehat{\mathcal{L}}_{X_i|Z_i}, \mathcal{L}_{X_i|Z_i}|D_2\right) \mathbb{E}_{\mathcal{L}}[\xi_i^2|Z_i, D_2] \xrightarrow{p} 0.$$

which is our assumption (19).

**Term** $B_n$ Again by Lemma 9 it is enough to show $\mathbb{E}[B_n^2 | Z, D_2] \xrightarrow{p} 0$. Using the fact that under the null conditionally on $Z, D_2$ the summands of $B_n$ are mean zero and independent we have that it is sufficient to show

$$\frac{1}{n\sigma_n^2}\sum \mathbb{E}[\xi_i^2|Z_i, D_2](\widehat{m}(Z_i) - m(Z_i))^2 \xrightarrow{p} 0 \tag{48}$$

Using the Lemma 11 we have

$$(\widehat{m}(Z_i) - m(Z_i))^2 \leqslant 2\left(1 + \chi^2\left(\widehat{\mathcal{L}}_{X_i|Z_i}, \mathcal{L}_{X_i|Z_i}|D_2\right)\right)\mathbb{E}[(\widehat{m}(X_i, Z_i) - m(Z_i))^2|Z_i, D_2]$$

using assumption (18) we have that (48) is implied by

$$\frac{1}{n\sigma_n^2}\sum \mathbb{E}[(\widehat{m}(X_i, Z_i) - m(Z_i))^2|Z_i, D_2]\mathbb{E}[\xi_i^2|Z_i, D_2] \xrightarrow{p} 0. \tag{49}$$

which is our assumption (20).

**Term** $C_n$ By Cauchy-Schwartz inequality we can upper bound $C_n$ by

$$\begin{aligned} C_n &\leqslant \frac{1}{\sqrt{n}\sigma_n}\left(\sum_{i=1}^n(\widehat{m}(Z_i) - m(Z_i))^2\right)^{1/2}\left(\sum_{i=1}^n(\widehat{m}(Z_i) - \mathbb{E}_{\mathcal{L}}[\widehat{m}(X_i, Z_i)|Z_i, D_2])^2\right)^{1/2} \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n(\widehat{m}(Z_i) - m(Z_i))^2\right)^{1/2}\left(\frac{1}{n\sigma_n^2}\sum_{i=1}^n(\widehat{m}(Z_i) - \mathbb{E}_{\mathcal{L}}[\widehat{m}(X_i, Z_i)|Z_i, D_2])^2\right)^{1/2}. \end{aligned}$$

Hence it is enough to show that

$$n\left(\frac{1}{n}\sum_{i=1}^n(\widehat{m}(Z_i) - m(Z_i))^2\right)\left(\frac{1}{n\sigma_n^2}\sum_{i=1}^n(\widehat{m}(Z_i) - \mathbb{E}_{\mathcal{L}}[\widehat{m}(X_i, Z_i)|Z_i, D_2])^2\right) = o_p(1) \tag{50}$$

Using Lemma 11 we conclude that the above display is implied by

$$\begin{aligned} &\left(\frac{1}{n}\sum_{i=1}^n\left(1 + \chi^2\left(\widehat{\mathcal{L}}_{X_i|Z_i}, \mathcal{L}_{X_i|Z_i}|D_2\right)\right)\mathbb{E}_{\mathcal{L}}[(\widehat{m}(X_i, Z_i) - m(Z_i))^2|Z_i, D_2]\right) \\ &\times\left(\frac{1}{n\sigma_n^2}\sum_{i=1}^n\chi^2\left(\widehat{\mathcal{L}}_{X_i|Z_i}, \mathcal{L}_{X_i|Z_i}|D_2\right)\mathbb{E}_{\mathcal{L}}[\xi_i^2|Z_i, D_2]\right) = o_p(n^{-1}) \end{aligned}$$

Under our assumption (18) it is sufficient to have

$$\left(\frac{1}{n}\sum\mathbb{E}_{\mathcal{L}}[(\widehat{m}(X_i, Z_i) - m(Z_i))^2|Z_i, D_2]\right)\times\left(\frac{1}{n\sigma_n^2}\sum_{i=1}^n\chi^2\left(\widehat{\mathcal{L}}_{X_i|Z_i}, \mathcal{L}_{X_i|Z_i}|D_2\right)\mathbb{E}_{\mathcal{L}}[\xi_i^2|Z_i, D_2]\right) = o_p(n^{-1})$$

which is our assumption (21).

Combining the convergence properties of the four terms, $T_N \xrightarrow{d} N(0, 1)$ by Slutsky's theorem. □

*Proof of Lemma 3.* Let us denote $u_i = m(Z_i) - \widehat{m}(Z_i)$ and $v_i = \mathbb{E}_{\mathcal{L}}(\widehat{m}(X_i, Z_i)|Z_i, D_2) - \widehat{m}(Z_i)$. Then we have that $R_i = (\varepsilon_i + u_i)(\xi_i + v_i)$. We have shown that $\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^n R_i \xrightarrow{d}$

$N(0,1)$ this implies $\frac{1}{n\sigma_n}\sum_{i=1}^n R_i \xrightarrow{p} 0$. Hence it is enough to show that $\frac{1}{n\sigma_n^2}\sum_{i=1}^n R_i^2 \xrightarrow{p} 1$,which would imply $T_D \xrightarrow{p} 1$. We decompose the term as

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^n R_i^2 = \overbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n \varepsilon_i^2\xi_i^2}^{S_1} + \overbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n v_i^2\varepsilon_i^2}^{S_2} + \overbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n u_i^2\xi_i^2}^{S_3} + \overbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n u_i^2 v_i^2}^{S_4}$$
$$+2\underbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n v_i\varepsilon_i^2\xi_i}_{C_1} + 2\underbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n u_i\varepsilon_i\xi_i^2}_{C_2} + 4\underbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n \varepsilon_i\xi_i u_i v_i}_{C_3}$$
$$2\underbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n u_i v_i^2\varepsilon_i}_{C_4} + 2\underbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n u_i^2 v_i\xi_i}_{C_5}$$

Let us look at one term at a time. We would show that all the terms except $S_1$ are $o_P(1)$ terms and $S_1 \xrightarrow{p} 1$. For showing $S_1 \xrightarrow{p} 1$ we invoke Lemma 8.

Observe that $\frac{1}{\sigma_n^2}\varepsilon_i^2\xi_i^2$ is an i.i.d sequence conditional on $D_2$ which mean 1. Hence if we assume $\sigma_n^{-(1+\delta)}\mathbb{E}\left(|\boldsymbol{\varepsilon\xi}|^{1+\delta} \mid \mathcal{F}_n\right) = o_P(n^\delta)$ (which is implied by the moment conditions needed for CLT a.k.a (22)) then we have that $S_1$ converges to 1 in probability.

We have that $S_2, S_3 = o_P(1)$ because $\mathbb{E}(S_2|Z, D_2) = \mathbb{E}(A_n^2|Z, D_2) = o_p(1)$ and $\mathbb{E}(S_3|Z, D_2) = \mathbb{E}(B_n^2|Z, D_2) = o_p(1)$ as already shown. For $S_4$ observe that

$$S_4 \leqslant \frac{1}{n\sigma_n^2}\sum_{i=1}^n u_i^2 \sum_{i=1}^n v_i^2 = o_p(1)$$

which is implied by (50), which we have already proved using (18) and (21). Next observe that

$$C_1 \leqslant \left(\frac{1}{n\sigma_n^2}\sum_{i=1}^n \varepsilon_i^2\xi_i^2\right)^{1/2}\left(\frac{1}{n\sigma_n^2}\sum_{i=1}^n v_i^2\varepsilon_i^2\right)^{1/2} = S_1^{1/2}S_2^{1/2} = o_p(1)$$
$$C_2 \leqslant S_1^{1/2}S_3^{1/2} \quad C_3 \leqslant S_3^{1/2}S_4^{1/2}$$
$$C_4 \leqslant S_4^{1/2}S_2^{1/2} \quad C_5 \leqslant S_4^{1/2}S_3^{1/2}$$

Since we have that $S_1 = O_p(1)$ and $S_i = o_p(1)$ for $i = 1, 2, 3$ we have that $C_k = o_p(1)$ for $k = 1, \ldots, 5$.

Combining everything so far we have that $\phi_n^{\text{tPCM}}$ is equivalent to the test: reject $H_0$ if

$$\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^n \varepsilon_i\xi_i \geqslant \frac{\widehat{\sigma}_n}{\sigma_n}z_{1-\alpha} - A_n - B_n - C_n$$

Now note that the RHS converges in probability to $z_{1-\alpha}$ and the oracle test statistic converges to $N(0,1)$ (hence does not accumulate near $z_{1-\alpha}$), hence by Lemma 10 we have that $\phi_n^{\text{vPCM}}$ is equivalent to $\phi_n^{\text{oracle}}$. □

*Proof of Lemma 1.* For our problem $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z}) \sim N(\boldsymbol{Z}^T\eta, 1)$ and $\widehat{\mathcal{L}}(\boldsymbol{X}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\hat{\eta}, 1)$. We also have that $m(\boldsymbol{X}, \boldsymbol{Z}) = \beta\boldsymbol{X} + \boldsymbol{Z}^T\gamma$ and $\widehat{m}(\boldsymbol{X}, \boldsymbol{Z}) = \hat{\beta}\boldsymbol{X} + \boldsymbol{Z}^T\hat{\gamma}$. Observe that $\xi_i = \widehat{m}(X_i, Z_i) - \mathbb{E}_{\mathcal{L}}(\widehat{m}(X_i, Z_i)|Z_i, D_2) = \hat{\beta}(X_i - \mathbb{E}(X_i|Z_i)) = \hat{\beta}\delta_i$.

Let us verify (16). Observe that $\mathrm{Var}(\varepsilon_i|Z_i, D_2) = 1$ and hence the required condition holds.

Next, we compute $\sigma_n^2$ as $\mathrm{Var}_{\mathcal{L}}[\xi_i|Z_i, D_2] = \hat{\beta}^2$, implying $\sigma_n^2 = \hat{\beta}^2$. We also evaluate the $\chi^2$ divergence between $\mathcal{L}_{\boldsymbol{X}|\boldsymbol{Z}}$ and $\widehat{\mathcal{L}}_{\boldsymbol{X}|\boldsymbol{Z}}$ using the identity (the identity can be verified by directly evaluating the divergence):

$$\chi^2(N(\mu, \sigma^2), N(\nu, \sigma^2)) = \exp\left(\frac{1}{\sigma^2}(\mu - \nu)^2\right) - 1$$

which yields $\chi^2(\mathcal{L}_{X_i|Z_i}, \widehat{\mathcal{L}}_{X_i|Z_i} \mid D_2) = \exp\left(\frac{1}{\sigma^2}[Z_i^T(\hat{\eta} - \eta)]^2\right) - 1$.

We observe that $\max_{i\in[n]} |Z_i^T(\hat{\eta} - \eta)| \leqslant \|Z_i\|_2\|\widehat{\eta} - \eta\|_2 \leqslant c_{\boldsymbol{Z}}\|\widehat{\eta} - \eta\|_2 \leqslant 1$ with high probability (since $\|\widehat{\eta} - \eta\|_2 \xrightarrow{p} 0$), hence on a high probability set:

$$\left(\exp\left(\frac{1}{\sigma^2}[Z_i^T(\hat{\eta} - \eta)]^2\right) - 1\right) \leqslant 2\frac{1}{\sigma^2}[Z_i^T(\hat{\eta} - \eta)]^2, \tag{51}$$

(where we used the fact that $e^x - 1 \leqslant 2x \,\forall\, 0 \leqslant x \leqslant 1$), this allows us to show (18) which follows using the property that $\max_{i\in[n]} |Z_i^T(\hat{\eta} - \eta)| \leqslant 1$ with high probability.

Let us look at the relevant error (19) which simplifies to

$$\frac{1}{n}\sum \chi^2(\mathcal{L}_{X_i|Z_i}, \widehat{\mathcal{L}}_{X_i|Z_i}) = \frac{1}{n}\sum\left(\exp\left(\frac{1}{\sigma^2}[Z_i^T(\hat{\eta} - \eta)]^2\right) - 1\right).$$

Using (51) it implies that it is enough to show that

$$\frac{1}{n}\sum_{i=1}^{n}(Z_i^T\hat{\eta} - Z_i^T\eta)^2 \xrightarrow{p} 0$$

which is clearly true because LHS is upper bounded by $c_{\boldsymbol{Z}}^2\|\widehat{\eta} - \eta\|_2^2$ which goes to zero in probability at a rate $\frac{1}{n}$. Let us look at the estimation error (20)

$$\mathbb{E}_{\mathcal{L}}[(\widehat{m}(X_i, Z_i) - m(Z_i))^2|Z_i, D_2] = \mathbb{E}_{\mathcal{L}}[(X_i\hat{\beta} + Z_i(\hat{\gamma} - \gamma))^2|Z_i, D_2].$$

It is enough to show that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{L}}[(X_i\hat{\beta} + Z_i(\hat{\gamma} - \gamma))^2|Z_i] \xrightarrow{p} 0$$

By further analysis, we obtain:

$$\begin{aligned}
\mathbb{E}_{\mathcal{L}}[(X_i\hat{\beta} + Z_i(\hat{\gamma} - \gamma))^2|Z_i] &= \mathbb{E}_{\mathcal{L}}[(Z_i^T\eta + \delta_i)\hat{\beta} + Z_i(\hat{\gamma} - \gamma))^2|Z_i] \\
&\leqslant 3\left(\widehat{\beta}^2\mathbb{E}[\delta_i^2|Z_i] + \widehat{\beta}^2(Z_i^T\eta)^2 + (Z_i^T(\widehat{\gamma} - \gamma))^2\right) \\
&\leqslant 3\left(\widehat{\beta}^2 + \widehat{\beta}^2 c_{\boldsymbol{Z}}^2\|\eta\|_2^2 + c_{\boldsymbol{Z}}^2\|\widehat{\gamma} - \gamma\|_2^2\right) = O_p\left(\frac{1}{n}\right)
\end{aligned}$$

where have used the fact that $\sqrt{n}\widehat{\beta} = O_p(1)$ and $\sqrt{n}\|\widehat{\gamma} - \gamma\|_2 = O_p(1)$. The last criterion (21) is product of the two rates going to zero at rate $1/n$ which is satisfed trivially because both the rates go to zero at rate $1/n$. □

## A.2 Proof of Results in Section 4.1

### A.2.1 Proof of main results

*Proof of Theorem 2.* $T_n^{\text{vPCM}}$ can be written as $T_N/T_D$ where $T_N = \frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^n L_i$ and $T_D = \widetilde{\sigma}_n/\sigma_n$ where $\widetilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n L_i^2 - \left(\frac{1}{n}\sum_{i=1}^n L_i\right)^2$. We would show that $T_N \xrightarrow{d} N(0,1)$ and $T_D \xrightarrow{p} 1$. First we make a crucial observation that

$$\begin{aligned}\widehat{f}(X_i,Z_i) - \mathbb{E}_{\mathcal{L}}[\widehat{f}(X_i,Z_i)|Z_i,D_2] &= \widehat{m}(X_i,Z_i) - \breve{m}(Z_i) - \mathbb{E}_{\mathcal{L}}\left[\widehat{m}(X_i,Z_i) - \breve{m}(Z_i)|Z_i,D_2\right] \\ &= \widehat{m}(X_i,Z_i) - \mathbb{E}_{\mathcal{L}}\left[\widehat{m}(X_i,Z_i)|Z_i,D_2\right] = \xi_i\end{aligned}$$

First we analyze $T_N$ for that we decompose $T_N$ into four terms as follows:

$$\begin{aligned}T_N = &\underbrace{\frac{1}{\sqrt{n}\sigma_n}\sum \varepsilon_i\xi_i}_{G_n'} + \underbrace{\frac{1}{\sqrt{n}\sigma_n}\sum \varepsilon_i(m_{\widehat{f}}(Z_i) - \widehat{m}_{\widehat{f}}(Z_i))}_{P_n} + \underbrace{\frac{1}{\sqrt{n}\sigma_n}\sum \xi_i(m(Z_i) - \widetilde{m}(Z_i))}_{Q_n} \\ &+ \underbrace{\frac{1}{\sqrt{n}\sigma_n}\sum (m_{\widehat{f}}(Z_i) - \widehat{m}_{\widehat{f}}(Z_i))(m(Z_i) - \widetilde{m}(Z_i))}_{R_n}\end{aligned}$$

We first focus on the term $G_n'$. We use Lemma S8 from (Lundborg et al., 2022), $\varepsilon_i\xi_i$ are conditionally independent given $\mathcal{F}_n \equiv \sigma(D_2)$. Also note that under the null conditional on $\mathcal{F}_n$, $\varepsilon_i\xi_i/\sigma_n$ are identically distributed random variables with mean zero and unit variance. Hence if we assume (assumption (22)) that

$$\frac{1}{\sigma_n^{2+\delta}}\mathbb{E}_{\mathcal{L}}\left[|\boldsymbol{\varepsilon}\boldsymbol{\xi}|^{2+\delta} \mid D_2\right] = o_P(n^{\delta/2})$$

we have that $G_n' \xrightarrow{d} N(0,1)$. Next we turn our attention to the term $P_n$. Our assumption (25) is equivalent to $\mathbb{E}_{\mathcal{L}}[P_n^2|Z,D_2] = o_p(1)$, now by using Lemma 9 we have that $P_n \xrightarrow{p} 0$. Similarly for the term $Q_n$ our assumption (24) is equivalent to $\mathbb{E}_{\mathcal{L}}[Q_n^2|Z,D_2] = o_p(1)$ which again by using Lemma 9 implies that $Q_n \xrightarrow{p} 0$. Finally we look at the fourth term $R_n$, by Cauchy-Schwartz inequality we can upper bound $R_n$ by

$$\begin{aligned}R_n &\leqslant \frac{1}{\sqrt{n}\sigma_n}\left(\sum_{i=1}^n (m_{\widehat{f}}(Z_i) - \widehat{m}_{\widehat{f}}(Z_i))^2\right)^{1/2}\left(\sum_{i=1}^n (m(Z_i) - \widetilde{m}(Z_i))^2\right)^{1/2} \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n (m_{\widehat{f}}(Z_i) - \widehat{m}_{\widehat{f}}(Z_i))^2\right)^{1/2}\left(\frac{1}{n\sigma_n^2}\sum_{i=1}^n (m(Z_i) - \widetilde{m}(Z_i))^2\right)^{1/2}.\end{aligned}$$

The RHS goes to zero in probability by our assumption (26) which implies $R_n = o_p(1)$.

Combining the convergence properties of the four terms, $T_N \xrightarrow{d} N(0,1)$ by Slutsky's theorem. Next we analyze $T_D$ and show it is $o_p(1)$.

Let us denote $u_i = m(Z_i) - \widetilde{m}(Z_i)$ and $v_i = m_{\widehat{f}}(Z_i) - \widehat{m}_{\widehat{f}}(Z_i)$. Then we have that $L_i = (\varepsilon_i + u_i)(\xi_i + v_i)$. We have shown that $\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^n L_i \xrightarrow{d} N(0,1)$ this implies $\frac{1}{n\sigma_n}\sum_{i=1}^n L_i \xrightarrow{p} 0$.

Hence it is enough to show that $\frac{1}{n\sigma_n^2}\sum_{i=1}^n L_i^2 \xrightarrow{p} 1$,which would imply $T_D \xrightarrow{p} 1$. We decompose the term as

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^n R_i^2 = \overbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n \varepsilon_i^2\xi_i^2}^{S_1} + \overbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n v_i^2\varepsilon_i^2}^{S_2} + \overbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n u_i^2\xi_i^2}^{S_3} + \overbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n u_i^2 v_i^2}^{S_4}$$
$$+2\underbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n v_i\varepsilon_i^2\xi_i}_{C_1} +2\underbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n u_i\varepsilon_i\xi_i^2}_{C_2} +4\underbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n \varepsilon_i\xi_i u_i v_i}_{C_3}$$
$$2\underbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n u_i v_i^2\varepsilon_i}_{C_4} +2\underbrace{\frac{1}{n\sigma_n^2}\sum_{i=1}^n u_i^2 v_i\xi_i}_{C_5}$$

Let us look at one term at a time. We would show that all the terms except $S_1$ are $o_P(1)$ terms and $S_1 \xrightarrow{p} 1$. For showing $S_1 \xrightarrow{p} 1$ we invoke Lemma S9 from (Lundborg et al., 2022). Observe that $\frac{1}{\sigma_n^2}\varepsilon_i^2\xi_i^2$ is an i.i.d sequence conditional on $D_2$ which mean 1. Hence if we assume $\sigma_n^{-(1+\delta)}\mathbb{E}\left(|\boldsymbol{\varepsilon\xi}|^{1+\delta} \mid \mathcal{F}_n\right) = o_P(n^\delta)$ (which is implied by the moment conditions needed for CLT a.k.a (22)) then we have that $S_1$ converges to 1 in probability.

We have that $S_2, S_3 = o_P(1)$ because $\mathbb{E}(S_2|Z, D_2) = \mathbb{E}(P_n^2|Z, D_2) = o_p(1)$ and $\mathbb{E}(S_3|Z, D_2) = \mathbb{E}(Q_n^2|Z, D_2) = o_p(1)$ as already shown. For $S_4$ observe that

$$S_4 \leqslant \frac{1}{n\sigma_n^2}\sum_{i=1}^n u_i^2 \sum_{i=1}^n v_i^2 = o_p(1)$$

which is implied by (26). Next observe that

$$C_1 \leqslant \left(\frac{1}{n\sigma_n^2}\sum_{i=1}^n \varepsilon_i^2\xi_i^2\right)^{1/2}\left(\frac{1}{n\sigma_n^2}\sum_{i=1}^n v_i^2\varepsilon_i^2\right)^{1/2} = S_1^{1/2}S_2^{1/2} = o_p(1)$$

$$C_2 \leqslant S_1^{1/2}S_3^{1/2} \quad C_3 \leqslant S_3^{1/2}S_4^{1/2}$$
$$C_4 \leqslant S_4^{1/2}S_2^{1/2} \quad C_5 \leqslant S_4^{1/2}S_3^{1/2}$$

Since we have that $S_1 = O_p(1)$ and $S_i = o_p(1)$ for $i = 1, 2, 3$ we have that $C_k = o_p(1)$ for $k = 1, \ldots, 5$.

Combining everything so far we have that $\phi_n^{\text{vPCM}}$ is equivalent to the test: reject $H_0$ if

$$\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^n \varepsilon_i\xi_i \geqslant \frac{\widetilde{\sigma}_n}{\sigma_n}z_{1-\alpha} - P_n - Q_n - R_n$$

Now note that the RHS converges in probability to $z_{1-\alpha}$ and the oracle test statistic converges to $N(0, 1)$ (hence does not accumulate near $z_{1-\alpha}$), hence by Lemma 10 we have that $\phi_n^{\text{vPCM}}$ is equivalent to $\phi_n^{\text{oracle}}$. □

## A.3 Proof of Results in Section 4.2

### A.3.1 Auxiliary Theorems and Lemmas

In this section we state a number of auxiliary lemmas and theorems which aid us in proving the main results. Many of them are borrowed from Niu et al. (2024) such as Lemma 13, 14 and 15, and Theorem 5, 6 and 7.

**Lemma 13** (Conditional convergence implies quantile convergence)**.** *Let $W_n$ be a sequence of random variables and $\alpha \in (0,1)$. If $W_n \mid \mathcal{F}_n \xrightarrow{d,p} W$ for some random variable $W$ whose CDF is continuous and strictly increasing at $\mathbb{Q}_\alpha[W]$, then*

$$\mathbb{Q}_\alpha\left[W_n \mid \mathcal{F}_n\right] \xrightarrow{p} \mathbb{Q}_\alpha[W].$$

**Lemma 14** (Conditional Jensen inequality)**.** *Let $W$ be a random variable and let $\phi$ be a convex function, such that $W$ and $\phi(W)$ are integrable. For any $\sigma$-algebra $\mathcal{F}$, we have the inequality*

$$\phi(\mathbb{E}[W \mid \mathcal{F}]) \leqslant \mathbb{E}[\phi(W) \mid \mathcal{F}] \text{ almost surely.}$$

**Lemma 15.** *Let $W_n$ be a sequence of random variables and $\mathcal{F}_n$ a sequence of $\sigma$-algebras. If $W_n \mid \mathcal{F}_n \xrightarrow{p,p} 0$, then $W_n \xrightarrow{p} 0$.*

**Theorem 4** (Conditional Slutsky's theorem)**.** *Let $W_n$ be a sequence of random variables. Suppose $a_n$ and $b_n$ are sequences of random variables such that $a_n \xrightarrow{p} 1$ and $b_n \xrightarrow{p} 0$. If $W_n \mid \mathcal{F}_n \xrightarrow{d,p} W$ for some random variable $W$ with continuous CDF, then*

$$a_n W_n + b_n \mid \mathcal{F}_n \xrightarrow{d,p} W$$

**Theorem 5** (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*

$$S_n^2 \equiv \sum_{i=1}^{n} \operatorname{Var}\left[W_{in} \mid \mathcal{F}_n\right],$$

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

$$\frac{1}{S_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\left[\left|W_{in} - \mathbb{E}\left[W_{in} \mid \mathcal{F}_n\right]\right|^{2+\delta} \mid \mathcal{F}_n\right] \xrightarrow{p} 0,$$

*then*

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

**Theorem 6** (Conditional law of large numbers)**.** *Let $W_{in}$ be a triangular array of random variables, such that $W_{in}$ are independent conditionally on $\mathcal{F}_n$ for each $n$. If for some $\delta > 0$ we have*

$$\frac{1}{n^{1+\delta}} \sum_{i=1}^{n} \mathbb{E}\left[\left|W_{in}\right|^{1+\delta} \mid \mathcal{F}_n\right] \xrightarrow{p} 0,$$

*then*

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

*The condition is satisfied when*

$$\sup_{1\leqslant i\leqslant n}\mathbb{E}\left[|W_{in}|^{1+\delta}\mid\mathcal{F}_n\right]=o_p\left(n^{\delta}\right).$$

**Theorem 7** (Unconditional weak law of large numbers)**.** *Let $W_{in}$ be a triangular array of random variables, such that $W_{in}$ are independent for each $n$. If for some $\delta>0$ we have*

$$\frac{1}{n^{1+\delta}}\sum_{i=1}^{n}\mathbb{E}\left[|W_{in}|^{1+\delta}\right]\to 0,$$

*then*

$$\frac{1}{n}\sum_{i=1}^{n}\left(W_{in}-\mathbb{E}\left[W_{in}\right]\right)\xrightarrow{p}0.$$

*The condition is satisfied when*

$$\sup_{1\leqslant i\leqslant n}\mathbb{E}\left[|W_{in}|^{1+\delta}\right]=o\left(n^{\delta}\right).$$

**Lemma 16.** *Under assumption* (22)

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{L}}(\varepsilon_i^2\mid Z_i)\mathbb{E}_{\mathcal{L}}(\xi_i^2\mid Z_i,D_2)\xrightarrow{p}1 \tag{52}$$

*and under assumption* (40)

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\left(\varepsilon_i^2-\mathbb{E}_{\mathcal{L}}(\varepsilon_i^2\mid Z_i)\right)\mathbb{E}(\widetilde{\xi}_i^2\mid Z_i,D_2)\xrightarrow{p}0 \tag{53}$$

*Under the previous two assumptions* (22), (40) *and additionally* (38) *we have that*

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\varepsilon_i^2\mathbb{E}(\widetilde{\xi}_i^2\mid Z_i)\xrightarrow{p}1 \tag{54}$$

*Proof.* We first show (52), let us define $W_{in}=\left(\mathbb{E}(\varepsilon_i^2\mid Z_i)\mathbb{E}(\xi_i^2\mid Z_i,D_2)\right)/\sigma_n^2$ and $\mathcal{F}_n=\sigma(D_2)$. Observe that $\mathbb{E}(W_{in}\mid D_2)=1$ since we are under the null. We will use Theorem 6 for which we need to bound the moments appropriately as follows:

$$\begin{aligned}
\frac{1}{n^{1+\delta/2}}\sum_{i=1}^{n}\mathbb{E}[|W_{in}|^{1+\delta/2}\mid\mathcal{F}_n] &= \frac{1}{n^{1+\delta/2}\sigma_n^{2+\delta}}\sum_{i=1}^{n}\mathbb{E}[|\left(\mathbb{E}(\varepsilon_i^2\mid Z_i)\mathbb{E}(\xi_i^2\mid Z_i,D_2)\right)|^{1+\delta/2}\mid D_2]\\
&= \frac{1}{n^{1+\delta/2}\sigma_n^{2+\delta}}\sum_{i=1}^{n}\mathbb{E}[|\left(\mathbb{E}(\varepsilon_i^2\xi_i^2\mid Z_i,D_2)\right)|^{1+\delta/2}\mid D_2]\\
&\leqslant \frac{1}{n^{1+\delta/2}\sigma_n^{2+\delta}}\sum_{i=1}^{n}\mathbb{E}[\mathbb{E}\left(|\varepsilon_i\xi_i|^{2+\delta}\mid Z_i,D_2\right)\mid D_2]\\
&= \frac{1}{n^{\delta}\sigma_n^{2+\delta}}\mathbb{E}[|\boldsymbol{\varepsilon\xi}|^{2+\delta}\mid D_2]=o_p(1)
\end{aligned}$$

The third line in the above display follows from Lemma 14 and the last line follows from assumption (22). Hence we have that

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{L}}(\varepsilon_i^2 \mid Z_i)\mathbb{E}_{\mathcal{L}}(\xi_i^2 \mid Z_i, D_2) \mid \mathcal{F}_n \xrightarrow{p,p} 1$$

from which (52) follows by applying Lemma 15.

Next we prove (53), let us define $W_{in} = \varepsilon_i^2\mathbb{E}(\widetilde{\xi}_i^2 \mid Z_i, D_2)/\sigma_n^2$ and $\mathcal{F}_n = \sigma(Z, D_2)$. Observe that $\mathbb{E}(W_{in} \mid \mathcal{F}_n) = \mathbb{E}(\varepsilon_i^2 \mid Z_i)\mathbb{E}(\widetilde{\xi}_i^2 \mid Z_i, D_2)/\sigma_n^2$. We use Theorem 6, for which we need to check some moment conditions:

$$\begin{aligned}
\frac{1}{n^{1+\delta/2}}\sum_{i=1}^{n}\mathbb{E}[|W_{in}|^{1+\delta/2} \mid \mathcal{F}_n] &= \frac{1}{n^{1+\delta/2}\sigma_n^{2+\delta}}\sum_{i=1}^{n}\mathbb{E}\left[\left|\left(\varepsilon_i^2\mathbb{E}(\widetilde{\xi}_i^2 \mid Z_i, D_2)\right)\right|^{1+\delta/2} \mid Z, D_2\right] \\
&\leqslant \frac{1}{n^{1+\delta/2}\sigma_n^{2+\delta}}\sum_{i=1}^{n}\mathbb{E}\left[|\varepsilon_i|^{2+\delta}\left(\mathbb{E}(|\widetilde{\xi}_i|^{2+\delta} \mid Z_i, D_2)\right) \mid Z, D_2\right] \\
&\leqslant \frac{1}{n^{1+\delta/2}\sigma_n^{2+\delta}}\sum_{i=1}^{n}\mathbb{E}\left[|\varepsilon_i|^{2+\delta} \mid Z_i, D_2\right]\mathbb{E}(|\widetilde{\xi}_i|^{2+\delta} \mid Z_i, D_2)
\end{aligned}$$

which goes to zero using our assumption (40). Hence we have by Theorem 6

$$\frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\left(\varepsilon_i^2 - \mathbb{E}_{\mathcal{L}}(\varepsilon_i^2 \mid Z_i)\right)\mathbb{E}(\widetilde{\xi}_i^2 \mid Z_i, D_2) \xrightarrow{p,p} 0.$$

which implies (53) by Lemma 15.

Next we will show (54):

$$\begin{aligned}
\frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\varepsilon_i^2\mathbb{E}(\widetilde{\xi}_i^2 \mid Z_i) = \frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\left(\varepsilon_i^2 - \mathbb{E}_{\mathcal{L}}(\varepsilon_i^2 \mid Z_i)\right)\mathbb{E}(\widetilde{\xi}_i^2 \mid Z_i, D_2) + \frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{L}}(\varepsilon_i^2 \mid Z_i)\mathbb{E}_{\mathcal{L}}(\xi_i^2 \mid Z_i, D_2) \\
+ \frac{1}{n\sigma_n^2}\sum_{i=1}^{n}\mathbb{E}(\varepsilon_i^2 \mid Z_i)\left[\mathbb{E}(\widetilde{\xi}_i^2 \mid Z_i) - \mathbb{E}(\xi_i^2 \mid Z_i)\right] \xrightarrow{p} 1
\end{aligned}$$

where we have used (52), (53) and (38). □

### A.3.2 Proof of the main results

*Proof of Lemma 2.* Observe that

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}(X_i, Z_i))^2 \leqslant C(Y, Z)$$

is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}(X_i, Z_i))^2 - \sum_{i=1}^{n}(Y_i - \mathbb{E}_{\widehat{\mathcal{L}}}[\widehat{m}(X_i, Z_i) \mid Z_i, D_2])^2 \leqslant \widetilde{C}(Y, Z)$$

where $\widetilde{C}(Y,Z)$ is the obtained by suitably updating $C(Y,Z)$. Now the above display can be shown equivalent to

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{\widehat{\mathcal{L}}}[\widehat{m}(X_i,Z_i)\mid Z_i,D_2]-\widehat{m}(X_i,Z_i)\right)\left(Y_i-\frac{\widehat{m}(X_i,Z_i)+\mathbb{E}_{\widehat{\mathcal{L}}}[\widehat{m}(X_i,Z_i)\mid Z_i,D_2]}{2}\right)\leqslant\widetilde{C}(Y,Z)\\
\iff\ &\frac{-2}{n}\sum_{i=1}^{n}(\widehat{m}(X_i,Z_i)-\widehat{m}(Z_i))(Y_i-\widehat{m}(Z_i))+\frac{1}{n}\sum_{i=1}^{n}(\widehat{m}(X_i,Z_i)-\widehat{m}(Z_i))^2\leqslant\widetilde{C}(Y,Z)\\
\iff\ &\frac{-2}{n}\sum_{i=1}^{n}(\widehat{m}(X_i,Z_i)-\widehat{m}(Z_i))(Y_i-\widehat{m}(Z_i))+\frac{1}{n}\sum_{i=1}^{n}((\widehat{m}(X_i,Z_i)-\widehat{m}(Z_i))^2-\mathbb{E}(\widetilde{\xi}_i^2\mid Z_i,D_2))\leqslant\widetilde{\widetilde{C}}(Y,Z)
\end{aligned}$$

We have adjusted by $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\xi_i^2\mid Z_i,D_2)$ on the last line and got the modified $\widetilde{\widetilde{C}}(Y,Z)$. Re-scaling by $-\frac{\sqrt{n}}{2\sigma_n}$ we have proved the result.

□

Note that Lemma 3 was proved in Section A.1.2.

*Proof of Lemma 4.* We have

$$\begin{aligned}
&\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^{n}(\widehat{\xi}_i^2-\mathbb{E}(\widetilde{\xi}_i^2\mid Z_i,D_2))\\
&\quad=\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^{n}\left(\widehat{\xi}_i^2-\mathbb{E}(\widehat{\xi}_i^2\mid Z_i,D_2)\right)\\
&\qquad+\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^{n}\left(\mathbb{E}(\widehat{\xi}_i^2\mid Z_i,D_2)-\mathbb{E}(\widetilde{\xi}_i^2\mid Z_i,D_2)\right)\\
&\quad=\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^{n}\left(\widehat{\xi}_i^2-\mathbb{E}(\widehat{\xi}_i^2\mid Z_i,D_2)\right)\\
&\qquad+\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^{n}(\mathbb{E}[\widehat{m}(X_i,Z_i)\mid Z_i,D_2]-\widehat{m}(Z_i))^2\\
&\qquad+\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^{n}(\mathbb{E}[\xi_i^2\mid Z_i,D_2]-\mathbb{E}[\widetilde{\xi}_i^2\mid Z_i,D_2])\\
&\quad\equiv I_n+II_n+III_n.
\end{aligned}$$

We have $I_n\xrightarrow{p}0$ because $\mathbb{E}[I_n^2\mid Z,D_2]$ by assumption (33). Furthermore, we have $II_n\xrightarrow{p}0$ and $III_n\xrightarrow{p}0$ by assumptions (34) and (35), respectively. □

*Proof of Lemma 5.* Observe that $T^{\mathrm{rHRT}}$ can be decomposed as

$$\begin{aligned}
&T^{\mathrm{rHRT}}(\widetilde{X},Y,Z)\\
&\quad=\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^{n}\varepsilon_i\widetilde{\xi}_i+\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^{n}(m(Z_i)-\widehat{m}(Z_i))\widetilde{\xi}_i-\frac{1}{2\sqrt{n}\sigma_n}\sum_{i=1}^{n}(\widetilde{\xi}_i^2-\mathbb{E}(\widetilde{\xi}_i^2\mid Z_i,D_2))\quad(55)\\
&\quad=I_n+II_n+III_n,
\end{aligned}$$

where $m(Z_i) = \mathbb{E}(Y_i \mid Z_i)$. First, we claim that $II_n, III_n \xrightarrow{p} 0$. Let us first look at $II_n$. We calculate

$$\mathbb{E}[II_n^2 \mid Z, D_2] = \frac{1}{n\sigma_n^2} \sum_{i=1}^{n} (m(Z_i) - \widehat{m}(Z_i))^2 \mathbb{E}(\widetilde{\xi}_i^2 \mid Z_i, D_2) \xrightarrow{p} 0.$$

The convergence to zero follows from our assumption (37), so we conclude that $II_n \xrightarrow{p} 0$ by Lemma 9. Next we look at $III_n$ and evaluate $\mathbb{E}(III_n^2 \mid D_2)$,which is given by

$$\mathbb{E}(III_n^2 \mid D_2) = \frac{\mathbb{E}\left[\mathrm{Var}(\widetilde{\boldsymbol{\xi}}^2 \mid \boldsymbol{Z}, D_2) \mid D_2\right]}{4\sigma_n^2} \xrightarrow{p} 0,$$

which goes to zero by our assumption (39), from which we conclude $III_n \xrightarrow{p} 0$ by Lemma 9. Now, we turn our attention to $I_n$. Let us denote by $W_{in} \equiv \varepsilon_i \widetilde{\xi}_i$, $\mathcal{F}_n = \sigma(Y, Z, D_2)$ and invoke the conditional CLT 5 to obtain that

$$\frac{1}{\widehat{S}_n} \sum_{i=1}^{n} W_{in} \xrightarrow{d,p} N(0,1), \tag{56}$$

where $\widehat{S}_n^2 = \sum_{i=1}^n \mathrm{Var}(W_{in} \mid \mathcal{F}_n) = \sum_{i=1}^n \varepsilon_i^2 \mathbb{E}(\widetilde{\xi}_i^2 \mid Z_i, D_2)$ if

$$\frac{1}{\widehat{S}_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}(|W_{in}|^{2+\delta} \mid \mathcal{F}_n) \xrightarrow{p} 0. \tag{57}$$

Now from Lemma 16 we know that $\frac{\widehat{S}_n^2}{n\sigma_n^2} \xrightarrow{p} 1$. Using this we know that (57) is equivalent to showing

$$\frac{1}{n^{1+\delta/2}\sigma_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}(|W_{in}|^{2+\delta} \mid \mathcal{F}_n) \xrightarrow{p} 0.$$

The LHS above is equal to

$$\begin{aligned}
\frac{1}{n^{1+\delta/2}\sigma_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}(|W_{in}|^{2+\delta} \mid \mathcal{F}_n) &= \frac{1}{n^{1+\delta/2}\sigma_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}(|\varepsilon_i \widetilde{\xi}_i|^{2+\delta} \mid Y_i, Z_i, D_2) \\
&= \frac{1}{n^{1+\delta/2}\sigma_n^{2+\delta}} \sum_{i=1}^{n} |\varepsilon_i|^{2+\delta} \mathbb{E}(|\widetilde{\xi}_i|^{2+\delta} \mid Z_i, D_2) \\
&\equiv IV_n.
\end{aligned}$$

Our assumption (40) implies $\mathbb{E}(IV_n \mid D_2) \xrightarrow{p} 0$, which by Lemma 9 implies $IV_n \xrightarrow{p} 0$. Hence the condition for the conditional CLT holds. Next let us look at the statement of conditional CLT, using the fact that $\frac{\widehat{S}_n^2}{n\sigma_n^2} \xrightarrow{p} 1$ we can show that (56) is equivalent to (by using conditional Slutsky, Theorem 4)

$$\frac{1}{\sqrt{n}\sigma_n} \sum_{i=1}^{n} \varepsilon_i \widetilde{\xi}_i \mid Y, Z, D_2 \xrightarrow{d,p} N(0,1).$$

Again using conditional Slutsky (Theorem 4) we have that $T^{\mathrm{rHRT}}(\widetilde{X}, Y, Z) \mid Y, Z, D_2 \xrightarrow{d,p} N(0,1)$. This in turn implies $C_n'(Y, Z) \xrightarrow{p} z_{1-\alpha}$ by Lemma 13. □

*Proof of Theorem 3.* By Lemma 2, we have

$$T_n^{\mathrm{HRT}} \geqslant C_n(Y,Z) \quad \Longleftrightarrow \quad T_n^{\mathrm{rHRT}} \geqslant C_n'(Y,Z). \tag{58}$$

By Lemmas 3, 4, and 5, we have

$$T_n^{\mathrm{rHRT}} \geqslant C_n'(Y,Z) \quad \Longleftrightarrow \quad T_n^{\mathrm{tPCM}} \geqslant C_n''(Y,Z), \quad \text{where} \quad C_n''(Y,Z) \xrightarrow{p} z_{1-\alpha}. \tag{59}$$

By Lemmas 3 and 12, we have $T_n^{\mathrm{tPCM}} \xrightarrow{d} N(0,1)$, so the non-accumulation condition (43) holds. Therefore, by Lemma 10, we conclude that the HRT and tPCM tests are asymptotically equivalent. □

*Proof of Lemma 6.* In Lemma 1 we have already verified all the assumptions pertaining to tPCM for the proposed linear model, so we only need to verify the assumptions (33), (34), (35), (37), (38), (39), and (40).

Recall from the proof of Lemma 1 that $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z}) \sim N(\boldsymbol{Z}^T\eta, 1)$ and $\widehat{\mathcal{L}}(\boldsymbol{X}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\widehat{\eta}, 1)$. We also have that $m(\boldsymbol{X},\boldsymbol{Z}) = \beta\boldsymbol{X} + \boldsymbol{Z}^T\gamma$ and $\widehat{m}(\boldsymbol{X},\boldsymbol{Z}) = \hat{\beta}\boldsymbol{X} + \boldsymbol{Z}^T\hat{\gamma}$. Observe that $\xi_i = \hat{\beta}\delta_i$, $\mathrm{Var}(\varepsilon_i|Z_i, D_2) = 1$ and $\mathrm{Var}_{\mathcal{L}}[\xi_i|Z_i, D_2] = \hat{\beta}^2$, implying $\sigma_n^2 = \hat{\beta}^2$. Now note that $\widehat{m}(Z_i) = \mathbb{E}_{\widehat{\mathcal{L}}}(\widehat{m}(X,Z)) = \widehat{\beta}(Z_i^T\widehat{\eta}) + Z_i^T\widehat{\gamma}$ and $\widetilde{\xi}_i = \widehat{m}(\widetilde{X}_i, Z_i) - \widehat{m}(Z_i) = \widehat{\beta}(\widetilde{X}_i - Z_i^T\widehat{\eta}) = \widehat{\beta}\delta_i'$, where $\delta_i' \overset{i.i.d}{\sim} N(0,1)$. This implies that $\mathbb{E}(\widetilde{\xi}_i^2 \mid Z_i, D_2) = \widehat{\beta}^2$.

First, we verify equation (33):

$$\begin{aligned}
&\frac{1}{\sigma_n^2}\mathbb{E}\left[\mathrm{Var}\left(\left(\widehat{m}(\boldsymbol{X},\boldsymbol{Z}) - \widehat{m}(\boldsymbol{Z})\right)^2 \mid \boldsymbol{Z}, D_2\right) \mid D_2\right] \\
&\quad = \frac{1}{\widehat{\beta}^2}\mathbb{E}\left[\mathrm{Var}\left(\widehat{\beta}^2\left(\boldsymbol{X} - \boldsymbol{Z}^T\widehat{\eta}\right)^2 \mid \boldsymbol{Z}, D_2\right) \mid D_2\right] \\
&\quad = \frac{\widehat{\beta}^4}{\widehat{\beta}^2}\mathbb{E}\left[\mathrm{Var}\left(\left(\boldsymbol{X} - \boldsymbol{Z}^T\widehat{\eta}\right)^2 \mid \boldsymbol{Z}, D_2\right) \mid D_2\right] \\
&\quad = \widehat{\beta}^2\mathbb{E}\left[\mathrm{Var}\left(\left(\boldsymbol{Z}^T(\eta - \widehat{\eta}) + \delta\right)^2 \mid \boldsymbol{Z}, D_2\right) \mid D_2\right] \\
&\quad \leqslant \widehat{\beta}^2\mathbb{E}\left[\mathbb{E}\left(\left(\boldsymbol{Z}^T(\eta - \widehat{\eta}) + \delta\right)^4 \mid \boldsymbol{Z}, D_2\right) \mid D_2\right] \\
&\quad \leqslant c_1\widehat{\beta}^2\mathbb{E}\left[\left(\boldsymbol{Z}^T(\eta - \widehat{\eta})\right)^4 + \delta^4 \mid D_2\right] \\
&\quad \leqslant c_1\widehat{\beta}^2\left[\|\widehat{\eta} - \eta\|_2^4 c_{\boldsymbol{Z}}^4 + \mathbb{E}\delta^4\right] = O_p\left(\frac{1}{n}\right)
\end{aligned}$$

where we have used the fact that $\widehat{\beta} = O_p\left(\frac{1}{\sqrt{n}}\right)$ and $\|\widehat{\eta} - \eta\|_2 = o_p(1)$.

Next, we verify equation (34):

$$\begin{aligned}
\frac{1}{\sqrt{n}\sigma_n}\sum_{i=1}^n(\widehat{m}(Z_i) - \mathbb{E}\left[\widehat{m}(X_i, Z_i) \mid Z_i, D_2\right])^2 &= \frac{1}{\sqrt{n}\widehat{\beta}}\sum_{i=1}^n\left(\widehat{\beta}(Z_i^T\widehat{\eta} - Z_i^T\eta)\right)^2 \\
&= \widehat{\beta}\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(Z_i^T\widehat{\eta} - Z_i^T\eta\right)^2 \\
&\leqslant |\sqrt{n}\widehat{\beta}||c_{\boldsymbol{Z}}^2\|\widehat{\eta} - \eta\|_2^2| = O_p\left(\frac{1}{n}\right),
\end{aligned}$$

where we have used Cauchy-Schwartz inequality at the last inequality and used the fact that $\widehat{\beta}$ and $\|\widehat{\eta} - \eta\|_2$ are $= O_p\left(\frac{1}{\sqrt{n}}\right)$.

Next, we note that equation (35) follows from the fact that $\mathrm{Var}_{\widehat{\mathcal{L}}}[\xi_i \mid Z_i, D_2] = \mathrm{Var}_{\mathcal{L}}[\xi_i \mid Z_i, D_2] = \widehat{\beta}^2$, which implies the LHS of (35) is exactly 0.

Next, we verify equation (37):

$$\begin{aligned}
\frac{1}{n\sigma_n^2}\sum_{i=1}^{n}(m(Z_i) - \widehat{m}(Z_i))^2\mathbb{E}(\widetilde{\xi}_i^2|Z_i, D_2) &= \frac{1}{n\widehat{\beta}^2}\sum_{i=1}^{n}\left[Z_i^T\gamma - (\widehat{\beta}(Z_i^T\widehat{\eta}) + Z_i^T\widehat{\gamma})\right]^2\widehat{\beta}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\beta}(Z_i^T\widehat{\eta}) + Z_i(\widehat{\gamma} - \gamma)\right]^2 \\
&\leqslant \widehat{\beta}^2\frac{1}{n}\sum_{i=1}^{n}(Z_i^T\widehat{\eta})^2 + \frac{1}{n}\sum_{i=1}^{n}(Z_i^T(\widehat{\gamma} - \gamma))^2 \\
&\leqslant \widehat{\beta}^2 c_{\boldsymbol{Z}}^2\|\widehat{\eta}\|_2^2 + \|\widehat{\gamma} - \gamma\|_2^2 c_{\boldsymbol{Z}}^2 = O_p\left(\frac{1}{n}\right)
\end{aligned}$$

The last line follows from the fact that $\widehat{\beta}^2$, $\|\widehat{\gamma} - \gamma\|_2^2$ and $\|\widehat{\eta} - \eta\|_2^2$ are $O_p\left(\frac{1}{n}\right)$.

To show (38), we observe that $\mathrm{Var}_{\widehat{\mathcal{L}}}[\xi_i \mid Z_i, D_2] = \mathrm{Var}_{\mathcal{L}}[\xi_i \mid Z_i, D_2] = \widehat{\beta}^2$ from which it follows that the the LHS is exactly zero, and hence (38) is trivially true.

Next, we verify equation (39):

$$\frac{\mathbb{E}\left[\mathrm{Var}(\widetilde{\boldsymbol{\xi}}^2 \mid \boldsymbol{Z}, D_2) \mid D_2\right]}{\sigma_n^2} = \frac{\widehat{\beta}^4}{\widehat{\beta}^2}\mathbb{E}(\mathrm{Var}(\delta'^2)) = \widehat{\beta}^2 = O_p\left(\frac{1}{n}\right)$$

Finally, we verify equation (40):

$$\frac{1}{\sigma_n^{2+\delta}}\mathbb{E}(|\boldsymbol{\varepsilon}\widetilde{\boldsymbol{\xi}}|^{2+\delta} \mid D_2) = \frac{1}{\widehat{\beta}^{2+\delta}}\widehat{\beta}^{2+\delta}\mathbb{E}(|\boldsymbol{\varepsilon}|^{2+\delta})\mathbb{E}(|\delta|^{2+\delta}) = O_p(1) = O_p(n^{\delta})$$

□

# B Method implementation details in the simulation study

**tPCM** We apply tPCM (Algorithm 3) with the `bam()` function from `mgcv` package for GAM fitting for $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ with penalization parameter `bs = "cs"`, and the banded precision matrix estimation from the `CovTools` package for $\mathcal{L}(\boldsymbol{X})$. We choose $B_{\text{tPCM}} = 25$ resamples and training proportion 0.4, the latter determined as described in Appendix D.

**HRT** We apply the HRT (Algorithm 2) with the `bam()` function from `mgcv` package for GAM fitting for $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ and the banded precision matrix estimation from the `CovTools` package for $\mathcal{L}(\boldsymbol{X})$. We choose $B_{\text{HRT}} = 5000$ resamples and training proportion 0.4. Because HRT was the slowest of the methods considered, we only applied it to the default simulation setting for the sake of computational feasibility.

**PCM** We apply a variant of PCM that is closer to Algorithm 1 from Lundborg et al. (2022) than vanilla PCM (Algorithm 1 in Section 2.1), as it includes Step 1 (ii) and Step 1 (iv). Step 1 (ii) was possible in this case since we fit a GAM. We continued to omit Step 2 of Algorithm 1 from Lundborg et al. (2022), which the authors claimed "is not critical for good power properties." We also use `bam()` for fitting $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}_{\text{-}j}]$. Moreover, to maintain a fair comparison, we endow PCM with knowledge of the banded covariance structure for the predictors. This meant that for any step where a function of $\boldsymbol{X}_j$ is regressed on $\boldsymbol{X}_{\text{-}j}$ (Steps 1 (iii) and 3 (i) from Algorithm 1 from Lundborg et al. (2022)), we actually only regressed $\boldsymbol{X}_j$ on $\boldsymbol{X}_{j-1}$ and $\boldsymbol{X}_{j+1}$, since $\boldsymbol{X}_j$ is independent of all other $\mathbf{X}_k$ given $\boldsymbol{X}_{j-1}$ and $\boldsymbol{X}_{j+1}$ under the banded structure. These regressions were also performed using `bam()`. We choose training proportion 0.3, determined as described in Appendix D.

**Oracle GCM** We also compare to an oracle version of the GCM test that is equipped with the true $\mathcal{L}(\boldsymbol{Y} \mid \boldsymbol{X})$ and $\mathcal{L}(\boldsymbol{X})$, as well as the same tower property-based acceleration as the tPCM test (also based on 25 resamples). Since there was no nuisance function estimation, there was no sample splitting, and so the Oracle GCM test had a larger sample size than the other methods.

# C Family-wise error rate in the simulation study

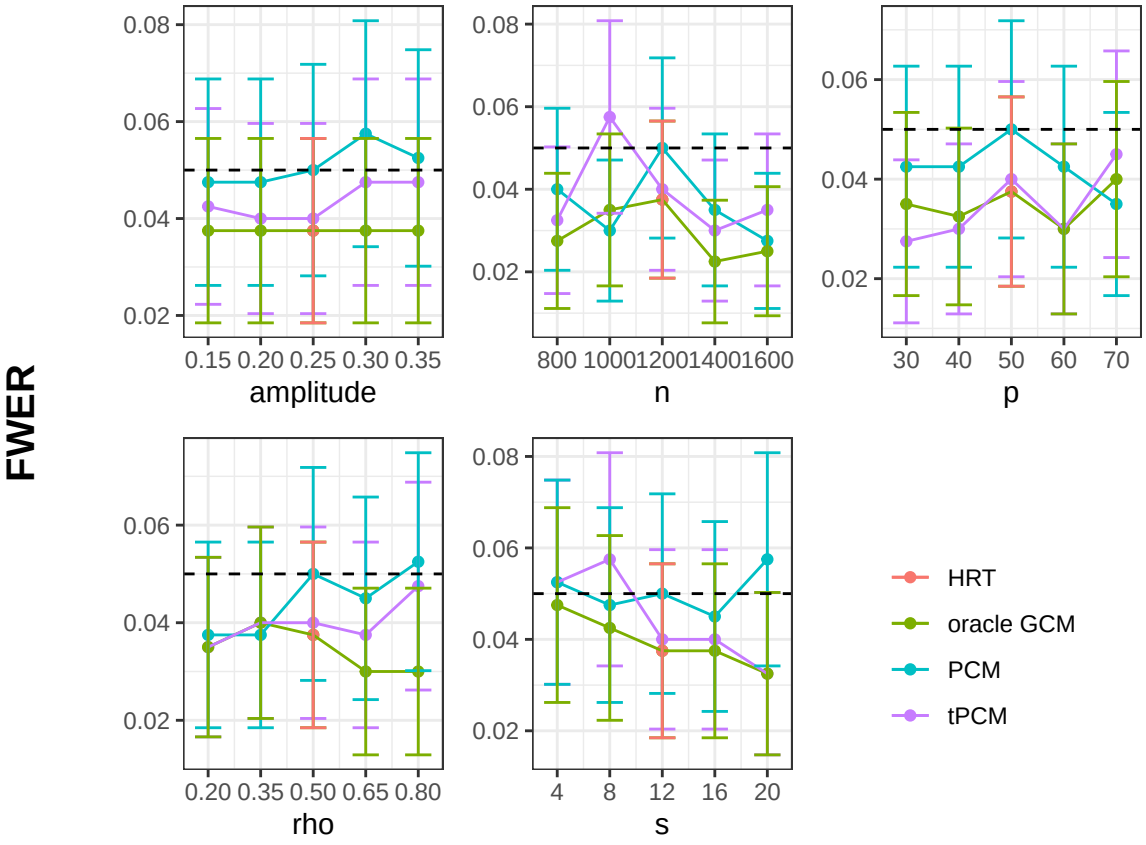


Figure 4: Type-I error control: in each plot, we vary one parameter. Each point is the average of 400 Monte Carlo replicates, and the error bars are the average $\pm 2 \times \widehat{\sigma}_f$, where $\widehat{\sigma}_f$ is the Monte Carlo standard deviation divided by $\sqrt{400}$.

# D Choosing the training proportions

In this section, we justify our choice of the best training proportions for tower PCM and PCM. For tPCM, we compared training proportions in $\{0.3, 0.4, 0.5, 0.6, 0.7\}$. For PCM we compared training proportions in $\{0.3, 0.4, 0.5\}$. We plot the family-wise error rates and power for for each method in Figures 6, 5, 7, and 8. In terms of type-I error for tPCM, 0.7 seems the most conservative which is perhaps not surprising, as it uses more data for the nuisances and less for testing. The rest of the proportions do not follow a monotonic trend, however. Generally, all proportions seem to be controlling the type-I error, though 0.5 and 0.6 exhibit some slight inflation for some settings. The type-I error rate for PCM is also not monotone. It is unclear what we should expect, since smaller training proportion means more data for the in-sample fits on the test split, but a poorer estimate of the direction of the alternative on the training split. In terms of power, though there is not a single training proportion that dominates uniformly for both tPCM and PCM, 0.4 and 0.3 are generally the highest, respectively.

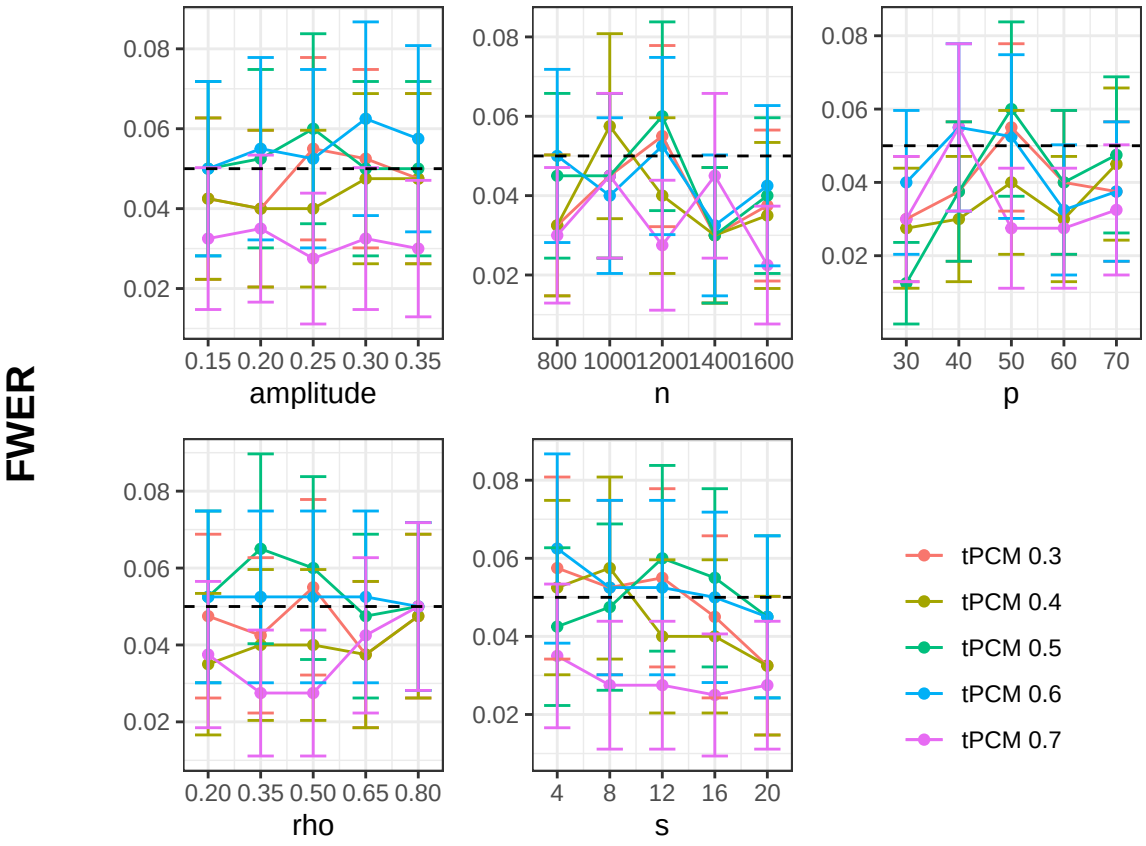


Figure 5: A family-wise error rate comparison of between different training proportions for tPCM.

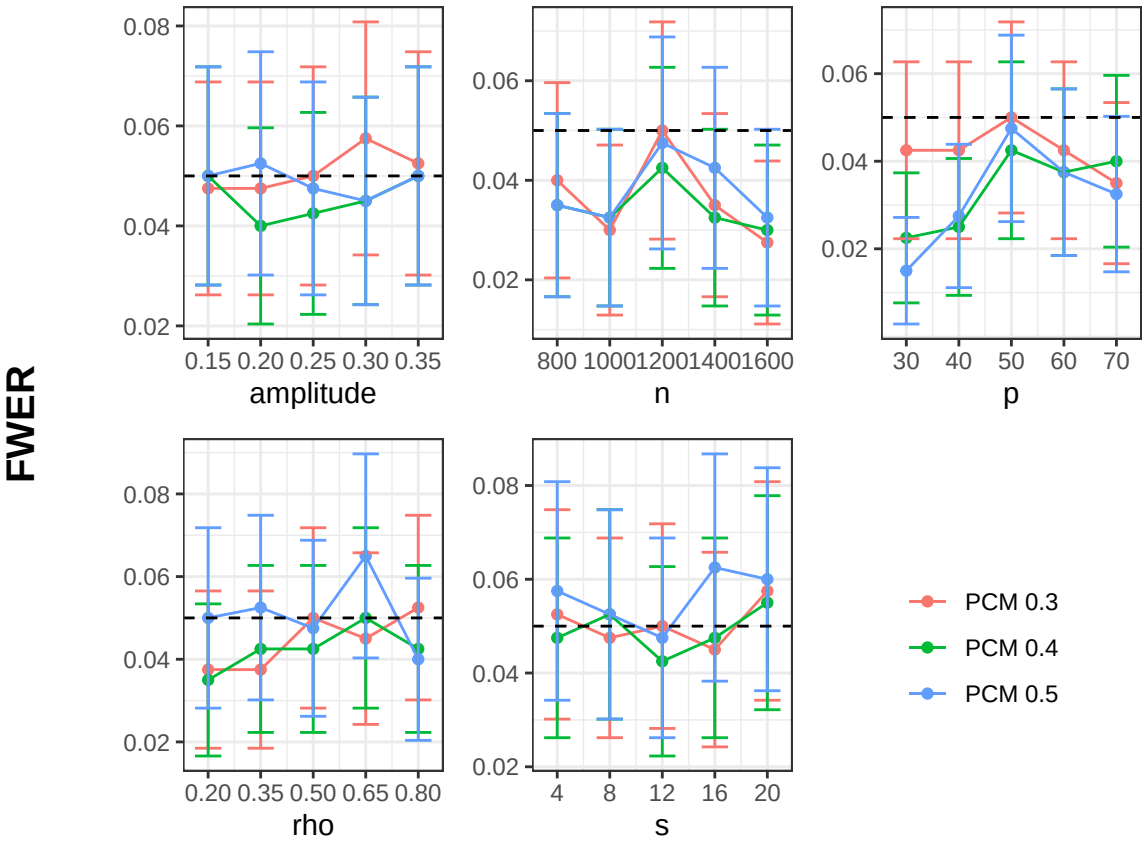


Figure 6: A family-wise error rate comparison between different training proportions for PCM.

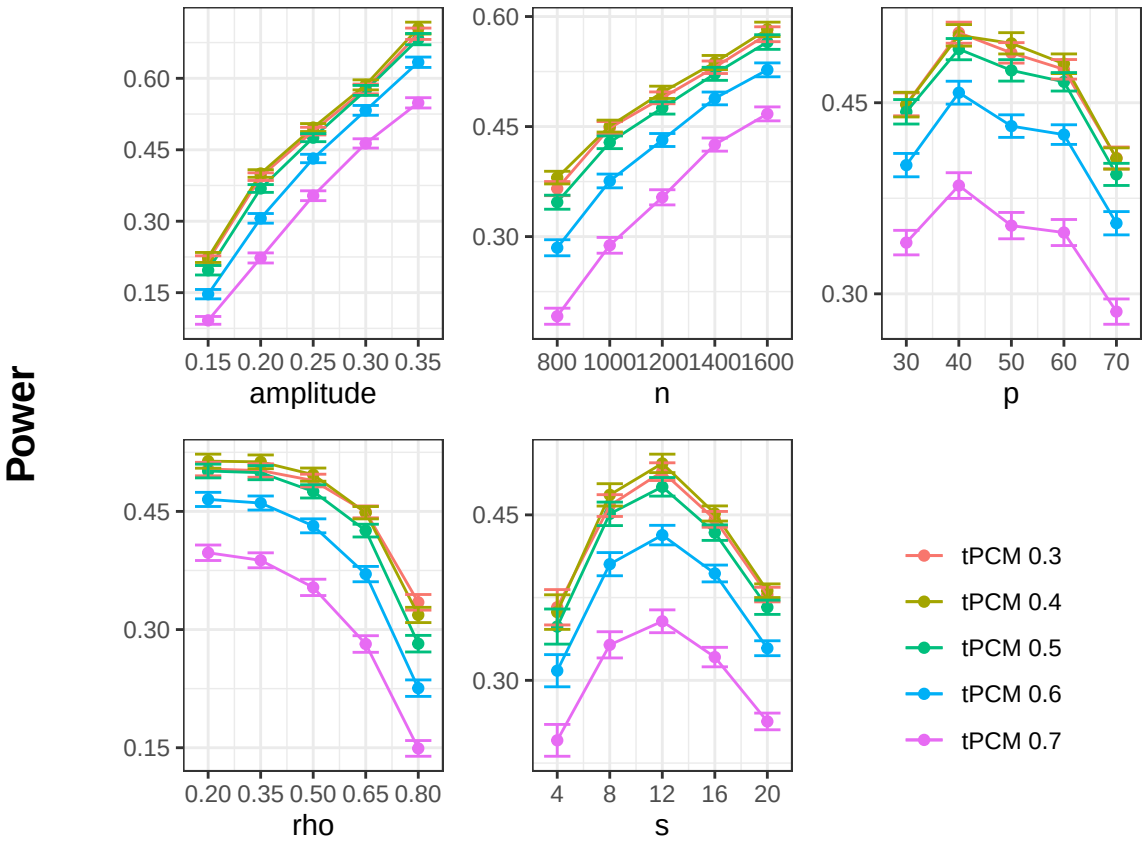


Figure 7: A power comparison between different training proportions for tPCM.

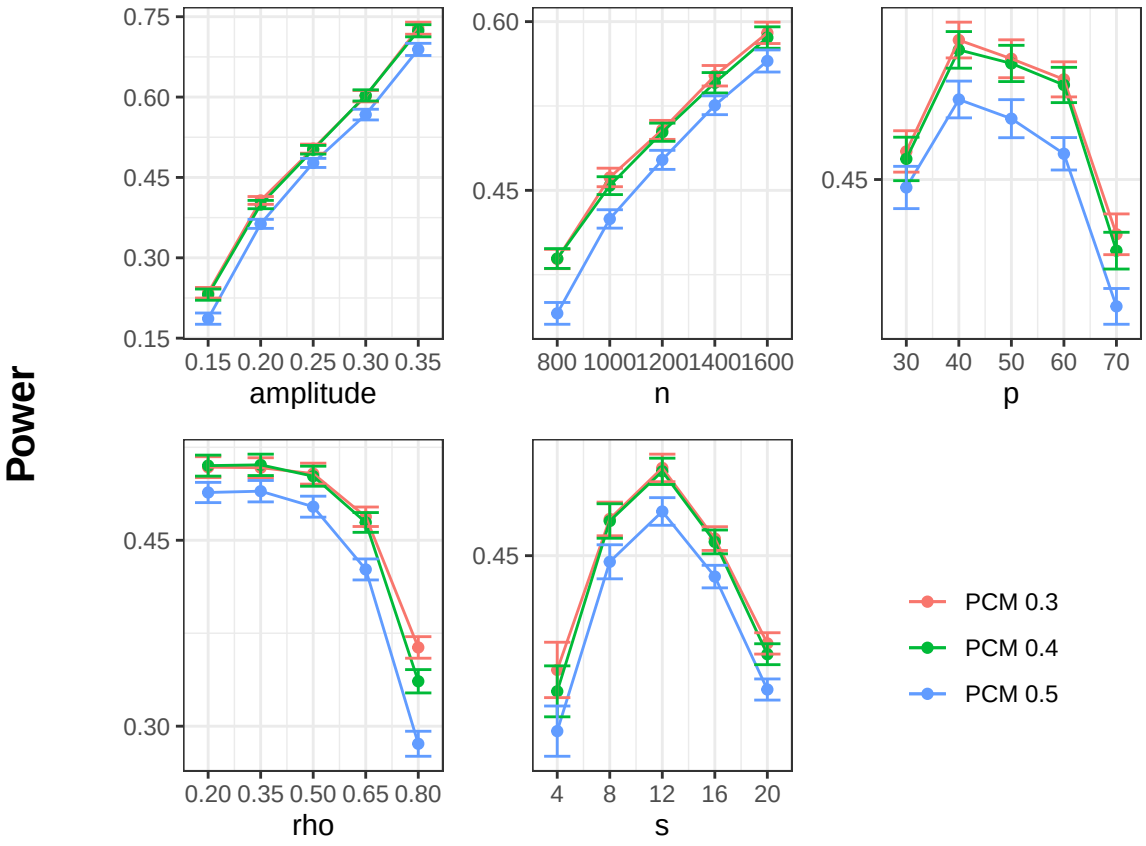


Figure 8: A power comparison between different training proportions for PCM.

# E Method implementation details in the data analysis

**HRT and tPCM** HRT and tPCM utilized a 0.35 training proportion. On the training sample, we obtained fits for $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ and $\mathcal{L}(\boldsymbol{X})$. For $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$, we used the same sparse GAM as implemented in the function `bam()` from `mgcv` with family = "gaussian" and penalization parameter `bs = "cs"`, as was done in our simulations. For $\mathcal{L}(\boldsymbol{X})$, we used the same fit as in Liu et al. (2022) and Li and Candès (2021), which was the graphical lasso as implemented in the `CVglasso()` function from the `CVglasso` package with parameter `lam.min.ratio = 1e-6`. HRT utilized $B_{\text{HRT}} = 5000 \approx 3 \times p/\alpha$ resamples, and tPCM used $B_{\text{tPCM}} = 25$ resamples to approximate conditional means.

**PCM** As in the simulation study, PCM was implemented as described in Algorithm 1 of Lundborg et al. (2022), except for Step 2, so it included the extra step (1(ii)) that can be performed when the contribution to $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ from $\boldsymbol{X}_j$ can be separated from the contributions from the other predictors, as is the case with a GAM. PCM also used a 0.35 sample split, and also used the sparse GAM implentation from the `bam()` function with family ="gaussian" as in the previous methods for fitting $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ on the training split, as well as for fitting $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}_{\text{-}j}]$ and $\mathbb{E}[f_j(\boldsymbol{X}_j) \mid \boldsymbol{X}_{\text{-}j}]$ on the evaluation split. A significant distinction from the simulation setting was that there is no banded structure with a known bandwidth of 1, so the regression of $f_j(\boldsymbol{X}_j)$ on $\boldsymbol{X}_{\text{-}j}$ had to include all of the $p-1$ predictors in $\boldsymbol{X}_{\text{-}j}$, rather than just $\boldsymbol{X}_{j-1}$ and $\boldsymbol{X}_{j+1}$.

**tGCM** tGCM is akin to the oracle GCM from the simulation, except $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ and $\mathcal{L}(\boldsymbol{X})$ are estimated from the data. tGCM uses the same tower-based acceleration as the tPCM test. There is no danger of a degenerate limiting distribution under the null, so we can make use of the full sample for testing through 5 fold cross-fitting. For each of the five equally sized folds, $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ and $\mathcal{L}(\boldsymbol{X})$ are estimated on the remaining 4/5 of the data using the same estimators as for HRT and tPCM. The tower trick is utilized to estimate $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}_{\text{-}j}]$ from the estimates for $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$ and $\mathcal{L}(\boldsymbol{X})$ using 25 resamples.