DIET: Conditional independence testing with marginal dependence measures of residual information

Mukund Sudarshan*[1] **Aahlad Puli*[1]** **Wesley Tansey[4]** **Rajesh Ranganath[1,2,3]**
[1]Computer Science, [2]Data Science, and [3]Population Health at Langone Health, New York University
[4]Computational Oncology Memorial Sloan Kettering Cancer Center

Abstract

Conditional randomization tests (CRTs) assess whether a variable $\mathbf{x}$ is predictive of another variable $\mathbf{y}$, having observed covariates $\mathbf{z}$. CRTs require fitting a large number of predictive models, which is often computationally intractable. Existing solutions to reduce the cost of CRTs typically split the dataset into a train and test portion, or rely on heuristics for interactions, both of which lead to a loss in power. We propose the decoupled independence test (DIET), an algorithm that avoids both of these issues by leveraging marginal independence statistics to test conditional independence relationships. DIET tests the marginal independence of two random variables: $F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z})$ and $F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z})$ where $F_{\cdot|\mathbf{z}}(\cdot \mid \mathbf{z})$ is a conditional cumulative distribution function (CDF) for the distribution $p(\cdot \mid \mathbf{z})$. These variables are termed "information residuals." We give sufficient conditions for DIET to achieve finite sample type-1 error control and power greater than the type-1 error rate. We then prove that when using the mutual information between the information residuals as a test statistic, DIET yields the most powerful conditionally valid test. Finally, we show DIET achieves higher power than other tractable CRTs on several synthetic and real benchmarks.

1 INTRODUCTION

A key question in many scientific disciplines is whether a variable $\mathbf{x}$ causes some outcome $\mathbf{y}$ (Lauritzen, 1996; Pearl, 2009). In genetics for example, scientists test whether a particular gene causes cancer to design targeted therapies (Zhu et al., 2018). When there are confounders $\mathbf{z}$ that may affect both $\mathbf{x}$ and $\mathbf{y}$, assessing the causal link between $\mathbf{x}$ and $\mathbf{y}$ corresponds to testing the conditional independence (CI) between $\mathbf{x}$ and $\mathbf{y}$ given $\mathbf{z}$:

$$\begin{aligned} \text{null hypothesis } \mathcal{H}_0 &: \mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z} \\ \text{alternate hypothesis } \mathcal{H}_1 &: \mathbf{x} \not\perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}. \end{aligned} \tag{1}$$

The advantage of using a hypothesis test for understanding such relationships is the ability to explicitly control the type-1 error rate: the probability of erroneously rejecting the null hypothesis where $\mathbf{x}$ is independent of $\mathbf{y}$ conditioned on $\mathbf{z}$. Consequently, constructing conditional independence hypothesis tests has become increasingly popular in the machine learning literature (Zhang et al., 2012; Doran et al., 2014; Sen et al., 2017; Runge, 2018; Bellot and van der Schaar, 2019).

Many existing tests however, have been shown to lose power when the dimensionality of $\mathbf{z}$ is high due to reliance on kernels (Bellot and van der Schaar, 2019) or fail to control the type-1 error rate when strong parametric assumptions about $p(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$ are violated (Candès et al., 2018).

To test for conditional independence when $\mathbf{z}$ is high-dimensional and without making assumptions on the form of $p(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$, Candès et al. (2018) proposed the conditional randomization test (CRT). The CRT calculates a p-value for eq. (1) by repeatedly comparing a scalar-valued test statistic $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}})$ with draws from the null distribution $T(\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)})$:

$$\frac{1}{M+1}\left(1 + \sum_{m=1}^{M} \mathbb{1}\left(T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}) \leq T(\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)})\right)\right), \tag{2}$$

where $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ is a set of N iid samples drawn from $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$. Null samples $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)}$ are drawn from the distribution $p(\mathbf{z}, \mathbf{y})p(\mathbf{x} \mid \mathbf{z})$, where $\widetilde{\mathbf{x}} \sim p(\mathbf{x} \mid \mathbf{z})$ is by construction conditionally independent of $\mathbf{y}$ given $\mathbf{z}$. If the null hypothesis is true, then $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}})$ will have the same distribution as each $T(\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)})$.

In contrast with other conditional independence testing methods, the CRT assumes the ability to sample $p(\mathbf{x} \mid \mathbf{z})$

*Equal contribution.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

but makes no assumptions on the form of $p(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$ or the test statistic T to control the type-1 error. This flexibility enables the use of powerful predictive models and empirical risk test statistics (Tansey et al., 2022; Liu et al., 2020; Sudarshan et al., 2021) that lead to higher power and better type-1 error rates than classical methods.

However, CRTs are computationally expensive. For each null sample, the test statistic must be recomputed. When using predictive models in empirical risk test statistics, these models must correspondingly be refit for *every* null sample $\mathcal{D}_{\tilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)}$. When the predictive models are computationally expensive to train, such as deep neural networks, the burden of running a CRT can become prohibitive.

Related work. There are two classes of conditional independence testing methods. These can be characterized by the assumptions they make to guarantee type-1 error control. We term the first class "Model-Y" methods, which make assumptions about the $\mathbf{y} \mid \mathbf{x}, \mathbf{z}$ distribution to ensure type-1 error control. This includes procedures that test for edges in Bayesian networks (Koller and Friedman, 2009; Spirtes et al., 2000; Cheng et al., 1998; De Campos and Huete, 2000), kernel-based methods (Fukumizu et al., 2007; Zhang et al., 2012), permutation-based methods (Gretton et al., 2012; Doran et al., 2014; Lee and Honavar, 2017), and many others.

The other class of conditional independence tests are "Model-X" methods, introduced by Candès et al. (2018). These require no assumptions about $\mathbf{y} \mid \mathbf{x}, \mathbf{z}$, but assume access to samples from $\mathbf{x} \mid \mathbf{z}$. This approach is more effective in controlling type-1 error than Model-Y methods when the number of labeled samples $(\mathbf{y}, \mathbf{x}, \mathbf{z})$ is small, but a large unlabeled dataset of $(\mathbf{x}, \mathbf{z})$ is available: like in the case of genetics. Much of the existing work on Model-X methods focuses on how to model $p(\mathbf{x} \mid \mathbf{z})$ from data (Romano et al., 2020; Sudarshan et al., 2020; Jordon et al., 2018), but leaves to the practitioner the form of the CRT test statistic $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}})$.

As a result, recent work in the Model-X space focuses on creating powerful but tractable CRT test statistics. Liu et al. (2020) propose a pair of methods called distilled conditional randomization tests (DCRTs). The first method, the d_0-CRT constructs a CRT where the test statistic is the marginal dependence between $(\mathbf{y} - \mathbb{E}[\mathbf{y} \mid \mathbf{z}])$ and $(\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{z}])$. However, Liu et al. (2020) demonstrate empirically that the d_0-CRT achieves low power when $\mathbf{y}$ is a function of some non-linear interaction between $\mathbf{x}$ and $\mathbf{z}$. To account for this issue, the authors also introduce the d_I-CRT. The d_I-CRT first uses a heuristic to select a small subset of $\mathbf{z}$ to explicitly construct a set of interaction terms with $\mathbf{x}$. It then fits a model $\hat{q}_{d_I}$ to estimate the conditional expectation of $\mathbf{y}$ given $(\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{z}]), \mathbb{E}[\mathbf{y} \mid \mathbf{z}]$, and each of the interaction terms. The d_I-CRT test statistic is some measure of feature importance of $\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{z}]$ in $\hat{q}_{d_I}$. If the heuristic pre-selection step fails to select the interactions that occur in the data, the d_I-CRT can fail to achieve power due to its reliance on conditional expectations.

The holdout randomization test (HRT) (Tansey et al., 2022) is another tractable yet flexible CRT. It splits samples of data into train and test sets, fits a predictive model on the train set, then uses this model to run a CRT only on the test set. While the HRT does not require heuristics for interactions between $\mathbf{x}$ and $\mathbf{z}$, it often loses power compared to DCRTs in practice due to sample splitting between the training and test set (Liu et al., 2020).

Patra et al. (2016) develop the notion of a nonparametric residual and study its use in testing for conditional independence, but do not provide a method with guarantees of power or type-1 error control. Residuals, like the ones computed in d_0-CRT and DIET, play a part in causal effect estimation under unobserved confounding. Objects called control functions are estimated as residuals from data and used to adjust for confounding: Guo and Small (2016) use additive residuals, Imbens and Newey (2009) use conditional CDFs like in DIET, and Puli and Ranganath (2020) give a general recipe to construct control functions with identification guarantees.

Our contributions. We propose a novel CRT to test $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$ that achieves high power without sample splitting. DIET first estimates two conditional CDFs: $F_{\mathbf{x}|\mathbf{z}}(\cdot \mid \cdot)$ and $F_{\mathbf{y}|\mathbf{z}}(\cdot \mid \cdot)$ using a dataset of samples from $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$. It then tests $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$ by testing the marginal independence of the univariate random variables produced by applying the conditional CDFs to $(\mathbf{x}, \mathbf{z})$ and $(\mathbf{y}, \mathbf{z})$ respectively: $F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z})$ and $F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z})$. DIET is computationally simple and, as we show, provides the ability to control type-I error regardless of the data generating distribution $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$. Further, we characterize distributions for which DIET can provably achieve power to correctly reject the null hypothesis.

Then, we discuss the limits of distillation procedures like DIET or the DCRTs: we highlight challenges a general procedure that distills a CI-test into a marginal one faces while maintaining type-I error control. After proving that further assumptions are necessary to overcome the challenges, we characterize conditions that allow one to reason about when a general distillation procedure provably achieves power. Finally, we validate DIET empirically on synthetic and real benchmarks and observe that it achieves higher power than several baselines while still controlling the type-1 error rate.

2 BACKGROUND

Conditional randomization tests (CRTs). CRTs outline a general procedure to test for the conditional independence of two variables $\mathbf{x}, \mathbf{y} \in \mathbb{R}$ given covariates $\mathbf{z} \in \mathbb{R}^p$. Us-

ing a dataset of N samples $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}} \in (\mathbb{R} \times \mathbb{R} \times \mathbb{R}^p)^N$ and a function $T : (\mathbb{R} \times \mathbb{R} \times \mathbb{R}^p)^N \to \mathbb{R}$, they compute the *test statistic* $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}})$. They then create *null datasets* ${\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}}^{(m)} \in (\mathbb{R}\times\mathbb{R}\times\mathbb{R}^p)^N$ by copying $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ and replacing $\mathbf{x}$ with new samples of $\widetilde{\mathbf{x}} \sim p(\mathbf{x} \mid \mathbf{z})$[1] to compute *null statistics* $T({\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}}^{(1)}), \ldots, T({\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}}^{(M)})$. Finally, CRTs use the true and null statistics to compute the p-value in eq. (2).

CRTs in the most general case compute $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}})$ by fitting and then evaluating the performance of a model $\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$ in predicting $\mathbf{y} \mid \mathbf{x}, \mathbf{z}$ (Tansey et al., 2022; Liu et al., 2020). To compute each null statistic, another model is fit and evaluated on each null dataset. $M + 1$ separate models must be fit because CRTs require that the same function T must be applied to both the true data and the null data. Given a user-specified false discovery rate (FDR) α and d CI tests, M is chosen to be $\mathcal{O}(\frac{d}{\alpha})$. For example, at a standard choice of $\alpha = 0.05$, and with just 100 variables, at least 2000 models need to be fit. This makes CRTs intractable.

In the next section, we introduce DIET: a flexible CRT that avoids sample splitting and heuristics like pre-selecting interaction terms.

3 DIET

Here we introduce a novel approach to distillation to create a tractable and powerful CRT. This section details the construction of the test statistic $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}})$, which measures the marginal dependence between $F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z})$ and $F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z})$. It then details the computation of each null statistic $T({\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}}^{(m)})$. Using the test and null statistics, DIET computes a p-value for testing $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$.

Fitting conditional CDF estimators. Let the CDFs associated with the distributions $p(\mathbf{x} \mid \mathbf{z})$ and $p(\mathbf{y} \mid \mathbf{z})$ be $F_{\mathbf{x}|\mathbf{z}}(\cdot \mid \cdot)$ and $F_{\mathbf{y}|\mathbf{z}}(\cdot \mid \cdot)$ respectively. DIET tests the marginal independence of the univariate random variables produced by applying the conditional CDFs to $(\mathbf{x}, \mathbf{z})$ and $(\mathbf{y}, \mathbf{z})$ respectively: $F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z})$ and $F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z})$. As a first step, DIET estimates these conditional CDFs with two estimators: $\hat{Q}_{\text{CDF}(\mathbf{x}|\mathbf{z})}(\cdot \mid \cdot\,; \theta)$ and $\hat{Q}_{\text{CDF}(\mathbf{y}|\mathbf{z})}(\cdot \mid \cdot\,; \eta)$. Any conditional CDF estimation technique can be used. Flexible examples include kernel-based methods (Bhattacharya and Gangopadhyay, 1990), nonparametric estimators, (Li and Racine, 2008), and mixture density networks (MDNs) (Bishop, 1994). We describe DIET with MDNs.

An MDN learns a neural network function $g : \mathbf{z} \mapsto \{\pi_\eta(\mathbf{z})[k], \mu_\eta(\mathbf{z})[k], \sigma_\eta(\mathbf{z})[k]\}_{k=1}^K$ to map values of $\mathbf{z}$ to the parameters of a gaussian mixture with K mixture components:

$$\hat{Q}_{\text{CDF}(\mathbf{y}|\mathbf{z})}(\mathbf{y} \mid \mathbf{z}; \eta) = \sum_{k=1}^{K} \pi_\eta(\mathbf{z})[k] \Phi\left(\frac{\mathbf{y} - \mu_\eta(\mathbf{z})[k]}{\sigma_\eta(\mathbf{z})[k]}\right).$$

The parameters η of $\hat{Q}_{\text{CDF}(\mathbf{y}|\mathbf{z})}(\mathbf{y} \mid \mathbf{z}; \eta)$ are learned via maximum likelihood estimation by optimizing over $(\mathbf{y}, \mathbf{z})$ pairs in dataset $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$:

$$\arg\max_{\eta} \frac{1}{N} \sum_{i=1}^{N} \log \hat{q}_{\text{PDF}}(\mathbf{y} = \mathbf{y}^{(i)} \mid \mathbf{z} = \mathbf{z}^{(i)}; \eta), \qquad (3)$$

where $\hat{q}_{\text{PDF}}$ is the conditional density implied by $\hat{Q}_{\text{CDF}}$. MDNs are useful as both the conditional CDF and density can be computed easily.

A model for $F_{\mathbf{x}|\mathbf{z}}(\cdot \mid \cdot)$, $\hat{Q}_{\text{CDF}(\mathbf{x}|\mathbf{z})}(\cdot \mid \cdot\,; \theta)$, is fit similarly but instead of using pairs of $(\mathbf{x}, \mathbf{z})$ from $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$, DIET uses only $\mathbf{z}$ from $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ and draw samples of $\widetilde{\mathbf{x}} \sim p(\mathbf{x} \mid \mathbf{z})$ for each $\mathbf{z}$ data point. Note that the distribution of $(\widetilde{\mathbf{x}}, \mathbf{z})$ is equal to that of $(\mathbf{x}, \mathbf{z})$, so evaluating $\hat{Q}_{\text{CDF}(\mathbf{x}|\mathbf{z})}(\cdot \mid \cdot\,; \theta)$ on samples of $(\mathbf{x}, \mathbf{z})$ from $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ will still be in-distribution.

Computing the test statistic $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}})$. The DIET test statistic measures the marginal dependence between two quantities $\hat{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{\delta}}$ using a dataset of paired samples $\mathcal{D}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}$. The variables $\hat{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{\delta}}$, termed "information residuals" represent the residual information contained in $\mathbf{x} \mid \mathbf{z}$ and $\mathbf{y} \mid \mathbf{z}$. They are computed as follows. A sample of $\hat{\boldsymbol{\epsilon}}$ is generated by evaluating the conditional CDF $\hat{Q}_{\text{CDF}(\mathbf{x}|\mathbf{z})}(\cdot \mid \cdot\,; \theta)$ at a sample $(\mathbf{x}, \mathbf{z})$, i.e. $\hat{\boldsymbol{\epsilon}} \leftarrow \hat{Q}_{\text{CDF}(\mathbf{x}|\mathbf{z})}(\mathbf{x} \mid \mathbf{z}\,; \theta)$. Similarly, $\hat{\boldsymbol{\delta}} \leftarrow \hat{Q}_{\text{CDF}(\mathbf{y}|\mathbf{z})}(\mathbf{y} \mid \mathbf{z}; \eta)$. To generate the dataset $\mathcal{D}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}$, a pair of $(\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\delta}})$ samples are computed for each $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ sample in $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ using the respective conditional CDFs.

Using the dataset of information residuals $\mathcal{D}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}$, DIET measures the marginal dependence between $\hat{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{\delta}}$ using the estimator of mutual information from Vinh et al. (2009). In practice, any measure of dependence $\rho : (\mathbb{R} \times \mathbb{R})^N \to \mathbb{R}$ can be used.

Computing null statistics $T({\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}}^{(m)})$. Computing each null statistic is very similar to computing the test statistic. First, a null dataset ${\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}}^{(m)}$ is sampled by copying $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$, then replacing the $\mathbf{x}$ values with $\widetilde{\mathbf{x}} \sim p(\mathbf{x} \mid \mathbf{z})$. The same $\hat{Q}_{\text{CDF}}$ models are used to generate information residuals using the null data, after which their mutual information is estimated. This process is repeated M times to generate M null statistics.

Computing a p-value. Using the test statistic $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}})$ and each null statistic $T({\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}}^{(m)})$, DIET computes a p-value using eq. (2). The full algorithm is summarized in algorithm 1.

While the DIET algorithm is relatively straightforward, it is not obvious why DIET should control the type-1 error

[1] All CRTs assume the ability to sample from $p(\mathbf{x} \mid \mathbf{z})$ to control type-1 error rates.

Algorithm 1: Decoupled independence test (DIET)

Input: Labeled dataset $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$, marginal dependence statistic ρ
Output: p-value $\hat{\mathbf{p}}$
Generate null dataset $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$ by replacing each $\mathbf{x}$ in $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ with a sample from $p(\mathbf{x} \mid \mathbf{z})$
Fit $\hat{Q}_{\text{CDF}(\mathbf{x}|\mathbf{z})}(\mathbf{x} \mid \mathbf{z}; \theta)$ and $\hat{Q}_{\text{CDF}(\mathbf{y}|\mathbf{z})}(\mathbf{y} \mid \mathbf{z}; \eta)$ using $(\mathbf{x}, \mathbf{z})$ pairs and $(\mathbf{y}, \mathbf{z})$ pairs from $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$
Generate null datasets $\{\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)}\}_{m=1}^{M}$
Create information residual dataset $\mathcal{D}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}$ by evaluating both $\hat{Q}_{\text{CDF}}$ models on $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$
for $m \in \{1, \ldots, M\}$ **do**
  Create null information residual dataset $\mathcal{D}^{(m)}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}$ by evaluating both $\hat{Q}_{\text{CDF}}$ models on $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)}$
end
$\hat{\mathbf{p}} \leftarrow \frac{1}{M+1}\left(1 + \sum_{m=1}^{M} \mathbb{1}\left[\rho(\mathcal{D}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}) \geq \rho(\mathcal{D}^{(m)}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}})\right]\right)$

rate, or achieve power. In the next section, we explore the theoretical properties of DIET.

4 THEORETICAL ANALYSIS OF DIET

Here we show that DIET achieves type-1 error control regardless of the data distribution. We then discuss when DIET can provably achieve power and characterize distributions where DIET is the most powerful test one can perform. The final part of this section provides a more general perspective on when distillation of a conditional randomization test into a marginal one is possible. We discuss how assumptions on the data generating process are always needed to guarantee power in a distillation procedure.

4.1 When can DIET control the type-1 error rate?

The type-1 error rate is the probability that the null hypothesis $\mathcal{H}_0$ is erroneously rejected: i.e. it is rejected when in reality $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$. To control this error rate at a user-specified level, the p-value under $\mathcal{H}_0$ must either be distributed uniformly over $[0, 1]$ or stochastically dominate[2] a Uniform$(0, 1)$ random variable (see appendix A.1 of Sudarshan et al. (2021) for a proof of this fact). Prop. 1 shows that DIET p-values computed using algorithm 1 will stochastically dominate a Uniform$(0, 1)$ random variable.

Proposition 1. *Let $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ be drawn from any distribution $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ and $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ consist of N iid samples from this distribution. If $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$, then for any measure of marginal dependence $\rho : (\mathbb{R} \times \mathbb{R})^N \to \mathbb{R}$ the* DIET *p-value computed using algorithm 1 will stochastically dominate a Uniform$(0, 1)$ random variable.*

[2] A random variable $\mathbf{a}$ stochastically dominates a random variable $\mathbf{b}$ if the following partial ordering exists on the CDFs of $\mathbf{a}$ and $\mathbf{b}$: $\forall x : F_{\mathbf{a}}(x) \leq F_{\mathbf{b}}(x)$.

We detail the full proof in appendix A.2, but provide a sketch here. Under $\mathcal{H}_0$, the test statistic $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}})$ is exchangeable with each of the null statistics $T(\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)})$. As a result, the p-value $\hat{\mathbf{p}}$ computed using eq. (2) will be uniformly distributed over the set $\{\frac{1}{M+1}, \frac{2}{M+1}, \ldots, 1\}$. Such a p-value stochastically dominates a Uniform(0, 1) random variable. Prop. 1 ensures that if the practitioner rejects the null hypothesis when $\hat{\mathbf{p}} \leq \alpha$, the probability of an erroneous rejection is no greater than the significance level α.

4.2 When can DIET provably achieve power?

A CRT achieves power when the distribution of $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}})$ is distinguishable from the distribution of each of the null statistics $T(\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)})$. Here we provide assumptions on the data distribution that will ensure that DIET is able to distinguish between the distribution of the test statistic versus the null statistics.

Theorem 1. *Let $F_{\cdot|\mathbf{z}}(\cdot \mid \mathbf{z})$ denote the conditional* CDF *for the distribution $p(\cdot \mid \mathbf{z})$. Let $\boldsymbol{\epsilon} = F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z})$ and $\boldsymbol{\delta} = F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z})$ be random variables defined over $(\mathbf{x}, \mathbf{z})$ and $(\mathbf{y}, \mathbf{z})$ respectively. Assume F is invertible in the first argument and $(\boldsymbol{\epsilon}, \boldsymbol{\delta}) \perp\!\!\!\perp \mathbf{z}$. If there exists a marginal independence test $\psi : (\mathbb{R} \times \mathbb{R})^N \times [0, 1] \to \{0, 1\}$ that uses a measure of dependence ρ and achieves power greater than $\alpha \in [0, 1]$, then* DIET *equipped with ρ and the conditional* CDFs *$F(\cdot \mid \mathbf{z})$ is a conditional independence test with power greater than α for data drawn from $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$.*

The conditional CDFs being invertible is a common assumption: e.g. when $\mathbf{x} \sim \mathcal{N}(\mathbf{z}_1, \sigma^2)$ or other continuous distributions. The core assumption here is that $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are jointly independent of the conditioning set of covariates $\mathbf{z}$. This independence $(\boldsymbol{\epsilon}, \boldsymbol{\delta}) \perp\!\!\!\perp \mathbf{z}$ holds in data generating processes where $\mathbf{x}, \mathbf{y}$ are strictly monotonic transformations of continuous noise variables for any fixed value fo $\mathbf{z}$; e.g. additive transformations like $\mathbf{x} = \mathbf{z} + \text{noise}$ and multiplicative transformations like $\mathbf{x} = \mathbf{z} * \text{noise}$. Appendix A.3.1 shows this formally.

We prove theorem 1 in appendix A.3: we show that given these conditions, DIET will provably be able to distinguish between the test and null statistics and achieve power to reject the null hypothesis. The proof establishes that when $\mathbf{x} \not\perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$, random variables $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ will be dependent. It also shows that under the null hypothesis $\mathcal{H}_0$, $\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\delta}$. Therefore, the test statistic, which measures the dependence of $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ will have a different distribution than the null statistics.

When is DIET the most powerful conditionally valid CRT? Here we show that under the same conditions as theorem 1, DIET equipped with a measure of mutual information ρ is the *most powerful* conditionally valid CRT (Katsevich and Ramdas, 2020).

The set of valid CRTs $\mathcal{C}_\alpha$ includes any CRT where the type-1 error is less than α using a dataset $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$. Given samples of $(\mathbf{y}, \mathbf{z})$, the set of conditionally valid CRTs at level α is a subset of $\mathcal{C}_\alpha$ where the samples of $(\mathbf{y}, \mathbf{z})$ in $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ are fixed. A conditionally valid CRT is also a marginally valid CRT. The following proposition states that given access to the conditional CDFs $F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z})$ and $F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z})$, DIET is the most powerful conditionally valid CRT. Thus, the power of DIET is tied directly to the quality of the estimation of these conditional CDFs.

Proposition 2. *Let* $\boldsymbol{\epsilon} = F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z})$ *and* $\boldsymbol{\delta} = F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z})$. *For data generating processes where both* $F_{\cdot|\mathbf{z}}(\cdot \mid \mathbf{z})$ *functions are invertible in the first argument and* $(\boldsymbol{\epsilon}, \boldsymbol{\delta}) \perp\!\!\!\perp \mathbf{z}$, DIET *with the following mutual information-based marginal dependence measure* ρ *is the most powerful conditionally valid test:*

$$\rho(\mathcal{D}_{\boldsymbol{\delta},\boldsymbol{\epsilon}}) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{p(\boldsymbol{\delta}_i, \boldsymbol{\epsilon}_i)}{p(\boldsymbol{\delta}_i)p(\boldsymbol{\epsilon}_i)}.$$

We prove prop. 2 in appendix A.5 by showing that the likelihood ratio in prop. 2 is equivalent to the likelihood ratio of $p(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$ and $p(\mathbf{y} \mid \mathbf{z})$: the most powerful conditionally valid CRT test statistic.

4.3 Multiple testing and variable selection

A common application of CRTs is controlled variable selection (Candès et al., 2018). Let $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_d\}$ be a set of covariates, and $\mathbf{y}$ be a response. Controlled variable selection methods identify a subset of important covariates by testing the conditional independence of each covariate $\mathbf{x}_j$ and $\mathbf{y}$ given all other covariates $\mathbf{x}_{-j}$. If the hypothesis test for $\mathbf{x}_j$ results in a rejection, that variable is "selected." The goal of controlled variable selection is to select as many variables as possible, while controlling for the FDR: an analog for type-1 error in multiple testing.

We apply the following procedure to use DIET for controlled variable selection (CVS). To test $\mathbf{x}_j \perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}_{-j}$ for each $\mathbf{x}_j$, we run algorithm 1 where $\mathbf{z} \leftarrow \mathbf{x}_{-j}$, $\mathbf{y} \leftarrow \mathbf{y}$, and $\mathbf{x} \leftarrow \mathbf{x}_j$. The resulting set of p-values is used with standard FDR-controlling procedures (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) to select important covariates.

4.4 Can we further generalize the assumptions made by DIET?

Is it possible to generalize the set of distributions for which power is achievable beyond DIET? We first outline what a general distillation procedure looks like using functions $u(\mathbf{x}, \mathbf{z})$ and $v(\mathbf{y}, \mathbf{z})$ to test for conditional independence $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$. If these functions u, v are to be learned using samples from $p(\mathbf{x} \mid \mathbf{z})p(\mathbf{y}, \mathbf{z})$ in order to provide type-I error control, we show the challenge faced by a general distillation procedure in always achieving power.

Limits of general distillation procedures. Let $L^2_{\mathbf{x},\mathbf{z}}$ denote the space of real-valued functions u of $(\mathbf{x}, \mathbf{z})$, where $\mathbb{E}[u(\mathbf{x}, \mathbf{z})^2] < \infty$. Let $L^2_{\mathbf{y},\mathbf{z}}$ be defined analogously. Rather than testing the marginal independence of conditional CDFs $F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z})$ and $F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z})$ like DIET, a *general distillation procedure* tests the marginal independence of some functions $u \in L^2_{\mathbf{x},\mathbf{z}}$ and $v \in L^2_{\mathbf{y},\mathbf{z}}$ instead. Daudin (1980) shows that, for all functions $u \in L^2_{\mathbf{x},\mathbf{z}}$ and $v \in L^2_{\mathbf{y},\mathbf{z}}$ such that $\mathbb{E}[u(\mathbf{x}, \mathbf{z}) \mid \mathbf{z}] = 0$ and $\mathbb{E}[v(\mathbf{y}, \mathbf{z}) \mid \mathbf{z}] = 0$,

$$\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z} \iff \mathbb{E}[u(\mathbf{x}, \mathbf{z})v(\mathbf{y}, \mathbf{z})] = 0.$$

This means that if $\mathbf{y}$ is conditionally *dependent* on $\mathbf{x}$ given $\mathbf{z}$, then there must exist functions u and v such that their correlation is non-zero. If these u and v are known beforehand, testing their marginal independence will yield a conditional independence test with power.

However, in reality u and v must be learned using data from the data distribution $p(\mathbf{y}, \mathbf{x}, \mathbf{z})$. As we show in corollary 2.1, using data from the *null-data-distribution*[3], $q_{null} = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z}, \mathbf{y})$, to learn u and v, guarantees type-I error control without the need to sample split or assume the functional form of $\mathbf{y} \mid \mathbf{x}, \mathbf{z}$, both of which lead to loss in power.

Learning from the null-data-distribution makes it hard to always achieve power. Consider the following data generating processes:

$$p_1(\mathbf{y}, \mathbf{x}, \mathbf{z}): \quad \mathbf{y} = \mathbf{x} + \mathbf{z} \mod 1 \qquad \mathbf{x}, \mathbf{z} \sim \text{Uniform}(0, 1)$$
$$p_2(\mathbf{y}, \mathbf{x}, \mathbf{z}): \quad \mathbf{y} = \mathbf{x} \qquad \mathbf{x}, \mathbf{z} \sim \text{Uniform}(0, 1),$$

where $\mathbf{a} + \mathbf{b} \mod 1$ is defined as $\mathbf{a} + \mathbf{b}$ if $\mathbf{a} + \mathbf{b} < 1$ and $\mathbf{a} + \mathbf{b} - 1$ if $\mathbf{a} + \mathbf{b} \geq 1$. Note that the marginals of $(\mathbf{x}, \mathbf{z})$ and $(\mathbf{y}, \mathbf{z})$ are the same across both p_1 and p_2. In turn, the null-data-distributions are equal, $p_1(\mathbf{x} \mid \mathbf{z})p_1(\mathbf{z}, \mathbf{y}) = p_2(\mathbf{x} \mid \mathbf{z})p_2(\mathbf{z}, \mathbf{y})$, meaning that any distillation procedure will learn the same functions u, v in either distribution. However, the same u, v can have dramatically different power in p_1 and p_2 making it difficult to build a generic distillation procedure. For example, let $u(\mathbf{x}, \mathbf{z}) = \mathbf{x} - 0.5$ and $v(\mathbf{y}, \mathbf{z}) = \mathbf{y} - 0.5$. Any general distillation procedure that tests the correlation between these u, v would yield power in p_2 but would have no power under p_1 because u, v are independent under p_1.

When do distillation procedures achieve power? Let $\boldsymbol{\epsilon} = u(\mathbf{x}, \mathbf{z}), \boldsymbol{\delta} = v(\mathbf{y}, \mathbf{z})$ be the variables computed by

[3] We use this name to denote that the null dataset $\mathcal{D}_{\tilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$, like in algorithm 1, is sampled from p_{null}

a distillation procedure, like DIET or d_0-CRT. The previous subsection explains the challenge any distillation procedure faces in both achieving power and having type-I error control. Here, we give conditions on the data generating process and the variables $\boldsymbol{\epsilon}, \boldsymbol{\delta}$ computed by the distillation procedure that guarantee power.

Theorem 2. *Consider any data generating process of the following form:*

$$\mathbf{z} \sim p(\mathbf{z}), \quad \mathbf{e}, \mathbf{d} \sim p(\mathbf{e}, \mathbf{d}), \quad \mathbf{x} = f(\mathbf{e}, \mathbf{z}) \quad \mathbf{y} = g(\mathbf{d}, \mathbf{z}).$$

Let $\boldsymbol{\epsilon}, \boldsymbol{\delta}$ be distributed according to:

$$(\boldsymbol{\epsilon}, \boldsymbol{\delta}, \mathbf{x}, \mathbf{y}, \mathbf{z}) \sim \hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) p(\mathbf{x}, \mathbf{y}, \mathbf{z}).$$

Further let,

$$\hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\boldsymbol{\epsilon} \mid \mathbf{x}, \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{y}, \mathbf{z}), \quad \text{(factorization)}$$

$$\exists \tilde{f}, \tilde{g} \qquad \mathbf{x} \overset{a.s.}{=} \tilde{f}(\boldsymbol{\epsilon}, \mathbf{z}), \quad \mathbf{y} \overset{a.s.}{=} \tilde{g}(\boldsymbol{\delta}, \mathbf{z}), \quad \text{(reconstruction)}$$

$$(\mathbf{d}, \boldsymbol{\delta}) \perp\!\!\!\perp \mathbf{z} \qquad (\mathbf{e}, \boldsymbol{\epsilon}) \perp\!\!\!\perp \mathbf{z}. \qquad \text{(joint independence)}$$

Let $\psi(\mathcal{D}_{\boldsymbol{\epsilon},\boldsymbol{\delta}}, \alpha) : (\mathbb{R} \times \mathbb{R})^N \times [0, 1] \to \{0, 1\}$ be a marginal independence test that uses statistic $\rho : (\mathbb{R} \times \mathbb{R})^N \to \mathbb{R}$ and has power greater than α. Let $\mathcal{D}_{\boldsymbol{\epsilon},\boldsymbol{\delta}}$ be a dataset of N samples of $(\boldsymbol{\epsilon}, \boldsymbol{\delta})$ generated using $\hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$ and $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$. Then, ψ using $\mathcal{D}_{\boldsymbol{\epsilon},\boldsymbol{\delta}}$ and ρ is also a conditional test of independence for $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$ with power greater than α.

We prove theorem 2 in appendix A.4. Theorem 2 allows one to use knowledge about the form of the data generating process to understand whether a distillation procedure achieves power. As an example, see appendix A.6 where we show how the d_0-CRT satisfies the conditions in theorem 2 for additive data generating processes and therefore achieves power for such processes.

5 EXPERIMENTS

We analyze the performance of DIET on several synthetic and real datasets and compare it to well-studied methods designed to make CRTs tractable.

DIET setup. The MDNs in DIET take $\mathbf{z}$ as input and use a six-layer fully-connected network with batch normalization and ReLU activations to output the parameters of a Gaussian mixture with 10 components. As a marginal dependence statistic ρ, we use the mutual information estimator from Vinh et al. (2009). Further training and hyperparameter details are given in appendix B.1.

Baselines. We use the d_0-CRT and d_1-CRT models described by Liu et al. (2020). The top-k $\mathbf{z}$ dimensions are chosen using the Lasso heuristic proposed by Liu et al. (2020). This model regresses $\mathbf{y}$ onto $\mathbf{z} \in \mathbb{R}^p$ and picks the top $k = 2 \log p$ dimensions of $\mathbf{z}$ with the largest absolute regression coefficients.

The HRTs we include in our experiments use a model $\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$ that consists of a six-layer fully-connected network with batch normalization and ReLU activations. We implement the cross-validated version of the HRT suggested by Tansey et al. (2022) that achieves higher power in finite samples.

Further details like the test statistics used for each baseline method can be found in appendix B.2.

Experiment details. Each synthetic experiment follows the same basic structure for a single run, unless specified otherwise. First, a dataset $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ is sampled. Then, each method is used to test the hypothesis $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$ and a p-value is computed using $M = 100$ null datasets. We perform 100 runs of each synthetic experiment and report aggregate results.

The power of each method at a specific rejection threshold α is estimated by computing the percentage of times a hypothesis is rejected, over the 100 runs. A hypothesis is rejected if the p-value $\hat{\mathbf{p}} \leq \alpha$.

For controlled variable selection experiments, we test the hypothesis $\mathbf{x}_j \perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}_{-j}$ for each dimension j of the covariate vector $\mathbf{x}$. We then apply the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to account for multiple testing while controlling the FDR.

To test each method in a realistic setting, the controlled variable selection experiments use only a fixed set of $\mathbf{x}$ samples. To generate the null datasets $\{\mathcal{D}_{\tilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)}\}_{m=1}^{M}$, we employ a deep generative model to jointly model each $p(\mathbf{x}_j \mid \mathbf{x}_{-j})$ distribution (Romano et al., 2020). Appendix B.3 provides an overview of this process. Since a deep generative model must be fit to generate null datasets, this experiment uses half the available data to fit the model, while the other half is used to run each CRT. Each synthetic variable selection experiment is run 100 times. We set $M = 2000$.

5.1 Synthetic experiments

Univariate Gaussian data. This experiment is designed mainly to confirm that each method performs as intended. The data is drawn as follows: $\mathbf{z} \sim \mathcal{N}(0, 0.1)$, $\mathbf{x} \mid \mathbf{z} \sim \mathcal{N}(\mathbf{z}, 0.1)$, and $\mathbf{y} \mid \mathbf{x}, \mathbf{z} \sim \mathcal{N}(\mathbf{x} + \mathbf{z}, 0.1)$. The training dataset consists of 500 samples.

Results: As expected, the estimated power of each method is 1 for $\alpha \in (0, 0.3]$. We do not explore larger α, as a practitioner would realistically set their nominal error rate within this range. As a graph is unnecessary to visualize this result, we omit it.

Non-Gaussian and multiplicative data. These experiments are designed primarily to understand the effect of violating an additivity assumption in the data generating

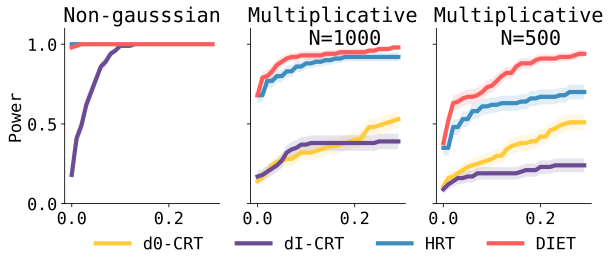

Figure 1: DIET achieves high power across numerous synthetic benchmarks. In this figure, we show the power of each method as a function of nominal type-1 error rate α.

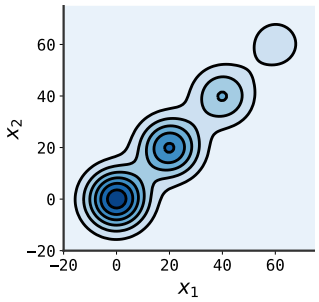

Figure 2: Synthetic CVS dataset

process. Using noise $\varepsilon \sim \mathcal{N}(0, 0.01)$ and coefficients $\beta \in \mathbb{R}^{100}$ where each $\beta_j \sim \mathcal{N}(0, 1)$ and sorted so that $|\beta_1| \geq |\beta_2| \geq \cdots, \geq \beta_d$, and $\mathbf{z} \sim \mathcal{N}(0, 0.01 \cdot I_{100})$

$$\mathbf{x} \mid \mathbf{z} \sim \mathcal{N}\left(\sum_{j=1}^{10} \beta_j \mathbf{z}_j, 0.25\right)$$

$$\mathbf{y} \mid \mathbf{x}, \mathbf{z}, \varepsilon = (\mathbf{x} + \varepsilon + \sum_{j=1}^{100} \mathbf{z}_j \beta_j)^3 \qquad \text{(Non-Gaussian)}$$

$$\mathbf{x} \sim \mathcal{N}(0, 1)$$

$$\mathbf{y} \mid \mathbf{z}, \mathbf{x}, \epsilon = 4\beta_1 \mathbf{z}_1 \mathbf{x} + 4\beta_2 \mathbf{z}_2 + \epsilon \qquad \text{(Multiplicative)}$$

Both datasets consist of 1000 samples.

Results: We observe that each CRT manages to control the type-1 error rate at or below nominal levels. In terms of power, most methods perform well on the non-Gaussian dataset, as shown in the first column of fig. 1. All but the d_I-CRT are able to achieve full power for almost every $\alpha \in (0, 0.3]$.

In the case of multiplicative data, there is a clear deterioration in the performance of the d_0-CRT and d_I-CRT, as shown in the second column of fig. 1. The d_I-CRT achieves marginally higher power for $\alpha < 0.2$, but is still quite far from DIET or HRT. Upon investigation, we observed that the heuristic used to choose dimensions in $\mathbf{z}$ in d_I-CRT only selects $\mathbf{z}_1$ at random. Since DCRTs forbid using samples of the triple $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ during training, it is difficult to choose a robust heuristic. We explore why DIET achieves higher power from a theoretical perspective in appendix A.1.

Then, to understand the cost of sample splitting, we reduced the sample size of the multiplicative data to 500 and re-ran our experiments. The third column of fig. 1 shows that the HRT suffers the greatest loss in power. This is likely due to the HRT splitting the sample and using only 200 samples during training.

Controlled variable selection. This experiment evaluates each CRT on its ability to perform controlled variable selection while using an estimated $p(\mathbf{x} \mid \mathbf{z})$ distribution. The $\mathbf{x}$ is a 100-dimensional mixture of autoregressive Gaussians. Figure 2 visualizes the first two dimensions of this data. The response $\mathbf{y} \mid \mathbf{x}$ is a conditional Gaussian whose mean is a linear function of $\mathbf{x}$ with only 20 non-zero coefficients. We refer the reader to appendix B.4 for the exact sampling process. The dataset consists of 1000 samples.

Results: We evaluate the average power and the false discovery proportion (FDP) across runs for each method in the fourth column of fig. 3 and fig. 4 respectively. The average FDP is an empirical estimate of the FDR. We notice that most methods are able to keep the average FDP below the nominal FDR rate α for $\alpha > 0.2$. However, when $\alpha \leq 0.1$, the d_I-CRT and the HRT inflate the FDP, suggesting they are sensitive to poor estimations of the $p(\mathbf{x}_j \mid \mathbf{x}_{-j})$ distributions, as shown by Sudarshan et al. (2021). We also observe that loss of power in the HRT is mainly due to sample splitting. Using 3000 samples instead helped increase the power of the HRT closer to that of DIET.

5.2 Semi-synthetic genetics experiment

A common application area of Model-X methods is biology (Candès et al., 2018; Bates et al., 2020; Sudarshan et al., 2020; Sesia et al., 2019). We evaluate each CRT using a setup similar to that of Sudarshan et al. (2020), which uses RNA expression data of 963 cancer cell lines and 20K genes per cell line from Yang et al. (2012). The datasets $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}} \in \mathbb{R}^{963 \times 100}$ are generated as follows.

100 genes are sampled sequentially from the set of 20K such that the resulting set contains genes with strong pairwise correlations. We use a synthetic $\mathbf{y} \mid \mathbf{x}$ response function from Tansey et al. (2022). Appendix B.5 contains specific details about the dataset creation. We perform 30 replicates of this experiment; $\mathbf{x}_{1:20}$ are the important features in each one.

Results: We show the average power for each CRT in the last column of fig. 3. All methods are able to control the average FDP below the nominal level. DIET consistently achieves power higher than the baselines. We also observe that the HRT achieves higher power than the d_I-CRT at nominal FDR above 0.1. At lower nominal FDR, the HRT does not select many features as its non-null p-values are generally higher than those of the d_I-CRT.

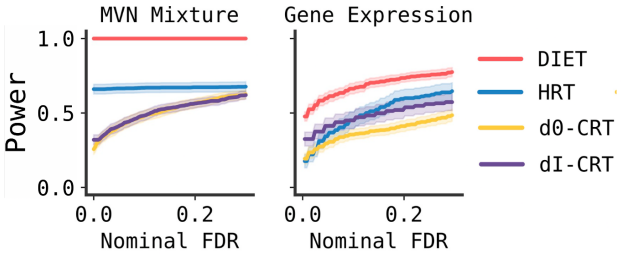

Figure 3: DIET achieves high power in CVS experiments. In this figure, we show the average power over 100 repetitions of each method as a function of nominal FDR in the case of variable selection.

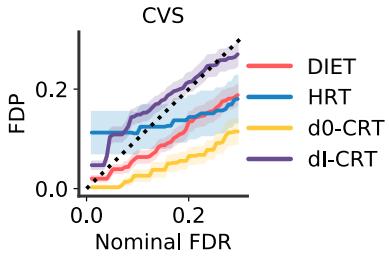

Figure 4: FDP of each method on synthetic CVS data.

5.3 Electronic health records

CRTs have found use in clinical model deployment pipelines as methods to prune a set of input features (Razavian et al., 2020). This pruning reduces the amount of auditing and engineering needed for model deployment. We perform controlled variable selection using an electronic health record (EHR) dataset from a large metropolitan hospital to understand which variables are most predictive of an adverse event within 96 hours for patients that tested positive for COVID-19.

The data contains 28K samples with 29 features on the results of a blood test, basic vital signs, and demographics. A full list of variables is provided in appendix B.6. We run each CRT method on the EHR dataset and apply the Benjamini and Hochberg (1995) procedure, selecting covariates at a nominal FDR of 10%.

Results: To evaluate the effectiveness of the selections made by each CRT, we compare selected covariates to those reported by several papers related to adverse events in COVID-19 patients from well-known medical journals (Petrilli et al., 2020; Sattar et al., 2020; Mei et al., 2020; Castro et al., 2020; Zhang et al., 2020; Zhong and Peng, 2021; Ruan et al., 2020; Zhou et al., 2020).

To score each CRT, we consider covariates found to be important by at least one of the above papers. We compute the fraction of these covariates selected by each CRT and report them in table 1. We show the full list of selections in appendix B.6.

DIET selects a larger percent of the important covariates, which indicates higher power. While the d_I-CRT selects almost as many, upon closer inspection, it also selects redundant features. For example, the d_I-CRT selects both count and percentage of Eosinophils, and both High O2 support and O2 device while DIET only selects one of each.

Table 1: DIET selects a larger portion of covariates previously identified by highly-cited medical papers. See appendix B.6 for a list of selections.

	DIET	HRT	d_0-CRT	d_I-CRT
Selected	**60%**	40%	25%	55%

6 DISCUSSION

Existing methods to speed up model-based CRTs either make restrictive assumptions about the data generating process, use heuristics to model interactions between $\mathbf{x}$ and $\mathbf{y}$, or lose power due to sample splitting. DIET provides a flexible way to avoid each of these issues and is applicable to a wide range of data generating distributions. It uses conditional CDF estimators to reduce high-dimensional model-based CRTs to tests of marginal independence.

We show theoretically that DIET will achieve type-1 error control regardless of data distribution $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$, then we characterize a class of data distributions for which DIET can provably achieve power. Future work in this area can study weaker assumptions on the data generating process to provably achieve power in a distillation-based CRT. This can lead to further insight into when a conditional independence test can be reduced to a marginal one without sacrificing power.

Acknowledgements

We thank the reviewers for their thoughtful comments. We would like to thank the participants of the Selective Inference Seminar for their helpful feedback: in particular Lihua Lei, Rina Barber, and Lucas Janson. This work was supported by the PhRMA Foundation Predoctoral Fellowship, NIH/NHLBI Award R01HL148248, NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science, NSF CAREER Award 2145542, NIH U54CA274492, R37CA271186, Break Through Cancer, and the Tow Center for Developmental Oncology.

References

S. Bates, M. Sesia, C. Sabatti, and E. Candès. Causal inference in genetic trio studies. *Proceedings of the National Academy of Sciences*, 117(39):24117–24126, 2020.

A. Bellot and M. van der Schaar. Conditional indepen-

dence testing using generative adversarial networks. *Advances in Neural Information Processing Systems*, 32: 2202–2211, 2019.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

P. K. Bhattacharya and A. K. Gangopadhyay. Kernel and nearest-neighbor estimation of a conditional quantile. *The Annals of Statistics*, pages 1400–1415, 1990.

C. M. Bishop. Mixture density networks. 1994.

E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3): 551–577, 2018.

V. M. Castro, T. H. McCoy, and R. H. Perlis. Laboratory findings associated with severe illness and mortality among hospitalized individuals with coronavirus disease 2019 in eastern massachusetts. *JAMA network open*, 3 (10):e2023934–e2023934, 2020.

J. Cheng, D. Bell, and W. Liu. Learning bayesian networks from data: An efficient approach based on information theory. *On World Wide Web at http://www. cs. ualberta. ca/˜ jcheng/bnpc. htm*, 1998.

J. Daudin. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590, 1980.

L. M. De Campos and J. F. Huete. A new approach for learning belief networks using independence criteria. *International Journal of Approximate Reasoning*, 24(1): 11–37, 2000.

G. Doran, K. Muandet, K. Zhang, and B. Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pages 132–141. Citeseer, 2014.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. *Advances in neural information processing systems*, 20, 2007.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Z. Guo and D. S. Small. Control function instrumental variable estimation of nonlinear causal effect models. *The Journal of Machine Learning Research*, 17(1): 3448–3482, 2016.

G. W. Imbens and W. K. Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009.

J. Jordon, J. Yoon, and M. van der Schaar. Knockoffgan: Generating knockoffs for feature selection using generative adversarial networks. In *International Conference on Learning Representations*, 2018.

E. Katsevich and A. Ramdas. A theoretical treatment of conditional independence testing under model-x. *arXiv preprint arXiv:2005.05506*, 2020.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. *arXiv preprint arXiv:1506.02557*, 2015.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

S. Lee and V. Honavar. A kernel conditional independence test for relational data. In *33rd Conference on Uncertainty in Artificial Intelligence, UAI 2017*, 2017.

Q. Li and J. S. Racine. Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics*, 26(4):423–434, 2008.

F. Liang, Q. Li, and L. Zhou. Bayesian neural networks for selection of drug sensitive genes. *Journal of the American Statistical Association*, 113(523):955–972, 2018.

M. Liu, E. Katsevich, L. Janson, and A. Ramdas. Fast and powerful conditional randomization testing via distillation. *arXiv preprint arXiv:2006.03980*, 2020.

Y. Mei, S. E. Weinberg, L. Zhao, A. Frink, C. Qi, A. Behdad, and P. Ji. Risk stratification of hospitalized covid-19 patients through comparative studies of laboratory results with influenza. *EClinicalMedicine*, 26:100475, 2020.

R. K. Patra, B. Sen, and G. J. Székely. On a nonparametric notion of residual and its applications. *Statistics & Probability Letters*, 109:208–213, 2016.

J. Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

C. M. Petrilli, S. A. Jones, J. Yang, H. Rajagopalan, L. O'Donnell, Y. Chernyak, K. A. Tobin, R. J. Cerfolio, F. Francois, and L. I. Horwitz. Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in new york city: prospective cohort study. *Bmj*, 369, 2020.

A. Puli and R. Ranganath. General control functions for causal effect estimation from ivs. *Advances in Neural Information Processing Systems*, 33, 2020.

N. Razavian, V. J. Major, M. Sudarshan, J. Burk-Rafel, P. Stella, H. Randhawa, S. Bilaloglu, J. Chen, V. Nguy, W. Wang, et al. A validated, real-time prediction model

for favorable outcomes in hospitalized covid-19 patients. *NPJ digital medicine*, 3(1):1–13, 2020.

Y. Romano, M. Sesia, and E. Candès. Deep knockoffs. *Journal of the American Statistical Association*, 115 (532):1861–1872, 2020.

Q. Ruan, K. Yang, W. Wang, L. Jiang, and J. Song. Clinical predictors of mortality due to covid-19 based on an analysis of data of 150 patients from wuhan, china. *Intensive care medicine*, 46(5):846–848, 2020.

J. Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947. PMLR, 2018.

N. Sattar, I. B. McInnes, and J. J. McMurray. Obesity is a risk factor for severe covid-19 infection: multiple potential mechanisms. *Circulation*, 142(1):4–6, 2020.

R. Sen, A. T. Suresh, K. Shanmugam, A. G. Dimakis, and S. Shakkottai. Model-powered conditional independence test. *Advances in neural information processing systems*, 30, 2017.

M. Sesia, C. Sabatti, and E. J. Candès. Gene hunting with hidden Markov model knockoffs. *Biometrika*, 106(1): 1–18, 2019.

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT press, 2000.

M. Sudarshan, W. Tansey, and R. Ranganath. Deep direct likelihood knockoffs. *Advances in Neural Information Processing Systems*, 33, 2020.

M. Sudarshan, A. Puli, L. Subramanian, S. Sankararaman, and R. Ranganath. Contra: Contrarian statistics for controlled variable selection. In *International Conference on Artificial Intelligence and Statistics*, pages 1900–1908. PMLR, 2021.

W. Tansey, V. Veitch, H. Zhang, R. Rabadan, and D. M. Blei. The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics*, 31(1):151–162, 2022.

P. K. Trivedi and D. M. Zimmer. *Copula modeling: an introduction for practitioners*. Now Publishers Inc, 2007.

N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080, 2009.

W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(D1):D955–D961, 2012.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

L. Zhang, X. Yan, Q. Fan, H. Liu, X. Liu, Z. Liu, and Z. Zhang. D-dimer levels on admission to predict in-hospital mortality in patients with covid-19. *Journal of Thrombosis and Haemostasis*, 18(6):1324–1329, 2020.

Q. Zhong and J. Peng. Mean platelet volume/platelet count ratio predicts severe pneumonia of covid-19. *Journal of clinical laboratory analysis*, 35(1):e23607, 2021.

F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, et al. Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. *The lancet*, 395(10229):1054–1062, 2020.

Z. Zhu, Z. Zheng, F. Zhang, Y. Wu, M. Trzaskowski, R. Maier, M. R. Robinson, J. J. McGrath, P. M. Visscher, N. R. Wray, et al. Causal associations between risk factors and common diseases inferred from gwas summary data. *Nature communications*, 9(1):1–12, 2018.

A APPENDIX

A.1 Shortcomings of DCRTs example

Consider the following example from earlier:

$$\begin{aligned}\mathbf{x} &\sim \mathcal{N}(\mathbf{x}; 0, \sigma_{\mathbf{x}}^2)\\ \mathbf{z}_j &\sim \mathcal{N}(\mathbf{z}_j; 0, 1)\, \forall j \in \{1, \ldots, d\}\\ \mathbf{y} \mid \mathbf{x}, \mathbf{z} &\sim \mathcal{N}(\mathbf{y}; \beta_1 \mathbf{x}\mathbf{z}_1 + \sum_{j=2}^{d} \beta_j \mathbf{z}_j, 1)\end{aligned}$$

Since this example extends the motivating example for d_I-CRTs from Liu et al. (2020), we focus only on the behavior of the d_I-CRT here. Recall the d_I-CRT test statistic computation:

1. The d_I-CRT first identifies a subset of k variables in $\mathbf{z}$ with which to explicitly compute interaction terms. This is done by fitting a regression from $\mathbf{z}$ to $\mathbf{y}$, then using some measure of feature importance to select the top k most important features, $\mathbf{z}_{\text{top}(k)}$
2. The distillation function $d_{\mathbf{y}} = \mathbb{E}[\mathbf{y} \mid \mathbf{z}]$ is computed
3. Then, the distillation function $d_{\mathbf{x}} = \mathbb{E}[\mathbf{x} \mid \mathbf{z}]$ is computed
4. Next, a model from $(\mathbf{x} - d_{\mathbf{x}}, d_{\mathbf{y}}, \mathbf{z}_{\text{top}(k)})$ to $\mathbf{y}$ is fit
5. Finally, a measure of feature importance for $\mathbf{x} - d_{\mathbf{x}}$ in this model is used to compute the test statistic T

To compute each null statistic, steps 3-5 are repeated using the null datasets. Given the set of M null statistics and the test statistic T, a p-value is computed as shown in the CRT p-value computation eq. (2). Now, observe the behavior of the d_I-CRT in this example.

First, a model is fit from $\mathbf{z}$ to $\mathbf{y}$. This is equivalent to estimating the function $\mathbb{E}[\mathbf{y} \mid \mathbf{z}]$. To see the functional form of this quantity let's first evaluate the density $F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z})$:

$$\begin{aligned}
F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z}) &= \int_{-\infty}^{\infty} f(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z}) d\mathbf{x}\\
&= \int_{-\infty}^{\infty} f(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) f(\mathbf{x}) d\mathbf{x}\\
&= \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-\frac{\left(-\beta_1 \mathbf{z}_1 \mathbf{x} - \sum_{j=2}^{d} \beta_j \mathbf{z}_j + \mathbf{y}\right)^2}{2} - \frac{\mathbf{x}^2}{2\sigma_{\mathbf{x}}{}^2}}}{\sqrt{4\pi^2 \sigma_{\mathbf{x}}{}^2}}\, \mathrm{d}\mathbf{x}\\
&= \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-\frac{\left(-M_1 \mathbf{x} + M_y\right)^2}{2} - \frac{\mathbf{x}^2}{2\sigma_{\mathbf{x}}{}^2}}}{\sqrt{4\pi^2 \sigma_{\mathbf{x}}{}^2}}\, \mathrm{d}\mathbf{x} \qquad \{\text{letting } M_1 = \beta_1 \mathbf{z}_1, M_y = \mathbf{y} - \sum_{j=2}^{d} \beta_j \mathbf{z}_j\}\\
&= \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-\frac{M_1^2 \mathbf{x}^2 - 2M_y M_1 \mathbf{x} + M_y^2}{2} - \frac{\mathbf{x}^2}{2\sigma_{\mathbf{x}}{}^2}}}{\sqrt{4\pi^2 \sigma_{\mathbf{x}}{}^2}}\, \mathrm{d}\mathbf{x}\\
&= \frac{\mathrm{e}^{-M_y^2/2}}{\sqrt{4\pi^2 \sigma_{\mathbf{x}}{}^2}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{M_1^2 \mathbf{x}^2 - 2M_y M_1 \mathbf{x}}{2} - \frac{\mathbf{x}^2}{2\sigma_{\mathbf{x}}{}^2}}\, \mathrm{d}\mathbf{x}\\
&= \frac{\mathrm{e}^{-M_y^2/2}}{\sqrt{4\pi^2 \sigma_{\mathbf{x}}{}^2}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\left(\mathbf{x}^2\left(M_1^2 + \frac{1}{\sigma_{\mathbf{x}}{}^2}\right) - 2M_y M_1 \mathbf{x} + \frac{M_y^2 M_1^2}{\left(M_1^2 + \frac{1}{\sigma_{\mathbf{x}}{}^2}\right)} - \frac{M_y^2 M_1^2}{\left(M_1^2 + \frac{1}{\sigma_{\mathbf{x}}{}^2}\right)}\right)}\, \mathrm{d}\mathbf{x}\\
&= \frac{\mathrm{e}^{-M_y^2/2}}{\sqrt{4\pi^2 \sigma_{\mathbf{x}}{}^2}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\left(\mathbf{x}\left(M_1^2 + \frac{1}{\sigma_{\mathbf{x}}{}^2}\right)^{1/2} - \frac{M_y M_1}{\left(M_1^2 + \frac{1}{\sigma_{\mathbf{x}}{}^2}\right)^{1/2}}\right)^2 + \frac{M_y^2 M_1^2}{\left(M_1^2 + \frac{1}{\sigma_{\mathbf{x}}{}^2}\right)}}\, \mathrm{d}\mathbf{x}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\mathrm{e}^{-M_y^2/2+\frac{M_y^2 M_1^2}{2\left(M_1^2+\frac{1}{{\sigma_{\mathbf{x}}}^2}\right)}}}{\sqrt{4\pi^2{\sigma_{\mathbf{x}}}^2}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{\left(M_1^2+\frac{1}{{\sigma_{\mathbf{x}}}^2}\right)}{2}\left(\mathbf{x}-\frac{M_y M_1}{\left(M_1^2+\frac{1}{{\sigma_{\mathbf{x}}}^2}\right)}\right)^2} \mathrm{d}\mathbf{x} \\
&= \frac{\mathrm{e}^{-M_y^2\left(\frac{1}{2}-\frac{1}{2\left(1+\frac{1}{M_1^2{\sigma_{\mathbf{x}}}^2}\right)}\right)}}{\sqrt{4\pi^2{\sigma_{\mathbf{x}}}^2}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{\left(M_1^2+\frac{1}{{\sigma_{\mathbf{x}}}^2}\right)}{2}\left(\mathbf{x}-\frac{M_y M_1}{\left(M_1^2+\frac{1}{{\sigma_{\mathbf{x}}}^2}\right)}\right)^2} \mathrm{d}\mathbf{x} \\
&= \frac{\mathrm{e}^{-M_y^2\left(\frac{1}{2}-\frac{1}{2\left(1+\frac{1}{M_1^2{\sigma_{\mathbf{x}}}^2}\right)}\right)}}{\sqrt{4\pi^2{\sigma_{\mathbf{x}}}^2}} \sqrt{\frac{2\pi}{\left(M_1^2+\frac{1}{{\sigma_{\mathbf{x}}}^2}\right)}} \\
&= \frac{\mathrm{e}^{-M_y^2\left(\frac{\left(1+\frac{1}{M_1^2{\sigma_{\mathbf{x}}}^2}\right)-1}{2\left(1+\frac{1}{M_1^2{\sigma_{\mathbf{x}}}^2}\right)}\right)}}{\sqrt{2\pi\left({\sigma_{\mathbf{x}}}^2 M_1^2+1\right)}} \\
&= \frac{\mathrm{e}^{-\frac{M_y^2}{2({\sigma_{\mathbf{x}}}^2 M_1^2+1)}}}{\sqrt{2\pi\left({\sigma_{\mathbf{x}}}^2 M_1^2+1\right)}} \\
&= \frac{\mathrm{e}^{-\frac{\left(\mathbf{y}-\sum_{j=2}^d \beta_j \mathbf{z}_j\right)^2}{2({\sigma_{\mathbf{x}}}^2 M_1^2+1)}}}{\sqrt{2\pi\left({\sigma_{\mathbf{x}}}^2 M_1^2+1\right)}} \\
&= \mathcal{N}(\mathbf{y}; \sum_{j=2}^{d} \beta_j \mathbf{z}_j, 1+\beta_1^2 \sigma_{\mathbf{x}}^2 \mathbf{z}_1^2).
\end{aligned}
$$

This is a Gaussian distribution with mean $\mathbb{E}[\mathbf{y} \mid \mathbf{z}] = \sum_{j=2}^d \beta_j \mathbf{z}_j$, which is not a function of $\mathbf{z}_1$. Therefore, $\mathbf{z}_{\text{top}(k)}$ will not include $\mathbf{z}_1$ for any $k < d$. To compute $d_{\mathbf{x}} = \mathbb{E}[\mathbf{x} \mid \mathbf{z}]$, note that $\mathbf{x}$ and $\mathbf{z}$ are independent, and $\mathbb{E}[\mathbf{x}] = 0$.

Next, let's consider a model from $(\mathbf{x} - d_{\mathbf{x}}, d_{\mathbf{y}}, \mathbf{z}_{\text{top}(k)})$ to $\mathbf{y}$. Again, this is equivalent to estimating $\mathbb{E}[\mathbf{y} \mid \mathbf{x} - d_{\mathbf{x}}, d_{\mathbf{y}}, \mathbf{z}_{\text{top}(k)}] = \mathbb{E}[\mathbf{y} \mid \mathbf{x}, \sum_{j=2}^d \beta_j \mathbf{z}_j, \mathbf{z}_{\text{top}(k)}]$. Since $\mathbf{z}_1$ is not in the conditioning set of this expectation, it reduces to $\mathbb{E}[\mathbf{y} \mid \sum_{j=2}^d \beta_j \mathbf{z}_j, \mathbf{z}_{\text{top}(k)}]$; this follows from expanding the conditional expectation and noting $\mathbb{E}[\mathbf{z}_1] = 0$. Thus any model from $(\mathbf{x} - d_{\mathbf{x}}, d_{\mathbf{y}}, \mathbf{z}_{\text{top}(k)})$ to $\mathbf{y}$ will assign no feature importance to $\mathbf{x} - d_{\mathbf{x}}$. Assuming that a feature importance score of 0 indicates an unimportant feature, the score assigned to $\mathbf{x} - d_{\mathbf{x}}$ will be 0.

The same holds true when repeating the d_I-CRT steps 3-5 with the null datasets. Regardless of what values of $\mathbf{x}$ are used in the model that estimates $\mathbb{E}[\mathbf{y} \mid \sum_{j=2}^d \beta_j \mathbf{z}_j, \mathbf{z}_{\text{top}(k)}]$, the importance score of $\mathbf{x} - d_{\mathbf{x}}$ will always be zero. Since the distribution of the test statistic is indistinguishable from the distribution of the null statistics, the d_I-CRT will achieve power no greater than the size of the test.

Next, consider the case of DIET. Recall that its test statistic uses the dataset $D_{\mathbf{x},\mathbf{y},\mathbf{z}} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^n$ to compute samples of $\boldsymbol{\delta} = F_{\mathbf{y}|\mathbf{z}}(\mathbf{y}, \mathbf{z})$ and $\boldsymbol{\epsilon} = F_{\mathbf{x}|\mathbf{z}}(\mathbf{x}, \mathbf{z}) = F_{\mathbf{x}}(\mathbf{x})$, then uses these samples to estimate the marginal dependence between $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$. We will now show that in the example above, $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ will be dependent using the true data $D_{\mathbf{x},\mathbf{y},\mathbf{z}}$, but will be independent when using the null data $D_{\tilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$, yielding power > 0.

First note the following equivalences:

$$
\begin{aligned}
F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z}) &= \Phi\left(\frac{\mathbf{y} - \sum_{j=2}^d \beta_j \mathbf{z}_j}{\sqrt{1+\beta_1^2 \sigma_{\mathbf{x}}^2 \mathbf{z}_1^2}}\right) \\
F_{\mathbf{x}}(\mathbf{x}) &= \Phi\left(\frac{\mathbf{x}}{\sigma_{\mathbf{x}}}\right)
\end{aligned}
$$

$$\mathbf{y} = \beta_1 \mathbf{x}\mathbf{z}_1 + \sum_{j=2}^{d} \beta_j \mathbf{z}_j + \eta_{\mathbf{y}}$$

where Φ is the CDF of a standard gaussian and $\eta_{\mathbf{y}} \sim \mathcal{N}(0,1)$. To show that $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ are dependent, we must show that

$$\mathbb{P}(\boldsymbol{\delta} \leq a \mid \boldsymbol{\epsilon} = b) \neq \mathbb{P}(\boldsymbol{\delta} \leq a).$$

When using the true data $D_{\mathbf{x},\mathbf{y},\mathbf{z}}$, the following must hold:

$$\begin{aligned}
\mathbb{P}(\boldsymbol{\delta} \leq a \mid \boldsymbol{\epsilon} = b) &= \mathbb{P}\left(\Phi\left(\frac{\mathbf{y} - \sum_{j=2}^{d} \beta_j \mathbf{z}_j}{\sqrt{1 + \beta_1^2 \sigma_{\mathbf{x}}^2 \mathbf{z}_1^2}}\right) \leq a \mid \Phi\left(\frac{\mathbf{x}}{\sigma_{\mathbf{x}}}\right) = b\right) \\
&= \mathbb{P}\left(\frac{\mathbf{y} - \sum_{j=2}^{d} \beta_j \mathbf{z}_j}{\sqrt{1 + \beta_1^2 \sigma_{\mathbf{x}}^2 \mathbf{z}_1^2}} \leq \Phi^{-1}(a) \mid \mathbf{x} = \sigma_{\mathbf{x}} \Phi^{-1}(b)\right) \\
&= \mathbb{P}\left(\frac{\beta_1 \mathbf{x}\mathbf{z}_1 + \eta_{\mathbf{y}}}{\sqrt{1 + \beta_1^2 \sigma_{\mathbf{x}}^2 \mathbf{z}_1^2}} \leq \Phi^{-1}(a) \mid \mathbf{x} = \sigma_{\mathbf{x}} \Phi^{-1}(b)\right) \\
&= \mathbb{P}\left(\frac{\beta_1 \mathbf{z}_1 \sigma_{\mathbf{x}} \Phi^{-1}(b) + \eta_{\mathbf{y}}}{\sqrt{1 + \beta_1^2 \sigma_{\mathbf{x}}^2 \mathbf{z}_1^2}} \leq \Phi^{-1}(a)\right).
\end{aligned}$$

The first equation uses the definitions of $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$. The second equation uses the invertibility of the Gaussian CDF. The third equation holds because $\mathbf{y}$ can be rewritten as a function of $\mathbf{x}$, $\mathbf{z}$, and noise $\eta_{\mathbf{y}}$. Finally, the last equation uses the value of $\mathbf{x}$ as a function of b and that $\mathbf{x}$ is jointly independent of $\mathbf{z}_1$ and $\eta_{\mathbf{y}}$. Clearly, the conditional probability $\mathbb{P}(\boldsymbol{\delta} \leq a \mid \boldsymbol{\epsilon} = b)$ cannot be written as $\mathbb{P}(\boldsymbol{\delta} \leq a)$ using the true data $D_{\mathbf{x},\mathbf{y},\mathbf{z}}$. This means that $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ will be dependent.

When computing the dependence of $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ using null datasets:

$$\begin{aligned}
\mathbb{P}(\boldsymbol{\delta} \leq a \mid \boldsymbol{\epsilon} = b) &= \mathbb{P}\left(\Phi\left(\frac{\mathbf{y} - \sum_{j=2}^{d} \beta_j \mathbf{z}_j}{\sqrt{1 + \beta_1^2 \sigma_{\mathbf{x}}^2 \mathbf{z}_1^2}}\right) \leq a \mid \Phi\left(\frac{\tilde{\mathbf{x}}}{\sigma_{\mathbf{x}}}\right) = b\right) \\
&= \mathbb{P}\left(\frac{\mathbf{y} - \sum_{j=2}^{d} \beta_j \mathbf{z}_j}{\sqrt{1 + \beta_1^2 \sigma_{\mathbf{x}}^2 \mathbf{z}_1^2}} \leq \Phi^{-1}(a) \mid \tilde{\mathbf{x}} = \sigma_{\mathbf{x}} \Phi^{-1}(b)\right) \\
&= \mathbb{P}\left(\frac{\beta_1 \mathbf{x}\mathbf{z}_1 + \eta_{\mathbf{y}}}{\sqrt{1 + \beta_1^2 \sigma_{\mathbf{x}}^2 \mathbf{z}_1^2}} \leq \Phi^{-1}(a) \mid \tilde{\mathbf{x}} = \sigma_{\mathbf{x}} \Phi^{-1}(b)\right) \\
&= \mathbb{P}\left(\frac{\beta_1 \mathbf{x}\mathbf{z}_1 + \eta_{\mathbf{y}}}{\sqrt{1 + \beta_1^2 \sigma_{\mathbf{x}}^2 \mathbf{z}_1^2}} \leq \Phi^{-1}(a)\right) \\
&= \mathbb{P}\left(\boldsymbol{\delta} \leq a\right).
\end{aligned}$$

The first 3 equations follow from earlier. The 4th and 5th steps hold because $\mathbf{y}$ is not a function of $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}$ is jointly independent of all other random variables. Therefore, when computing each null statistic using null data $D_{\tilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$, $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ will be independent.

Since DIET-CDF will identify dependence between $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ when using the true data, and no dependence when using the null data, the distribution of the test statistic will not be equal to that of each null statistic. Thus, it follows that DIET-CDF can achieve power > 0.

A.2 Proof of prop. 1

Proposition 1. *Let $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ be drawn from any distribution $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ and $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ consist of N iid samples from this distribution. If $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$, then for any measure of marginal dependence $\rho : (\mathbb{R} \times \mathbb{R})^N \to \mathbb{R}$ the* DIET *p-value computed using algorithm 1 will stochastically dominate a Uniform$(0,1)$ random variable.*

Proof. Recall the DIET p-value introduced in algorithm 1:

$$\hat{\mathbf{p}} = \frac{1}{M+1}\left(1 + \sum_{m=1}^{M} \mathbb{1}\left[\rho(\mathcal{D}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}) \geq \rho(\mathcal{D}^{(m)}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}})\right]\right).$$

We will prove that if a $\hat{q}$ estimator is trained on data $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$, the above p-value will be super-uniform. Using the technique from Candès et al. (2018), it suffices to show that the following sequence is exchangeable under the null, conditional on samples of $(\mathbf{z}, \mathbf{y})$:

$$\rho(\mathcal{D}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}), \rho(\mathcal{D}^{(1)}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}), \ldots, \rho(\mathcal{D}^{(M)}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}).$$

Note that $\mathcal{D}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}$, and $\{(\mathcal{D}^{(m)}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}})\}_{m=1}^{M}$ are datasets of information residuals. As such, the above sequence can be rewritten as:

$$\rho(\{\hat{\boldsymbol{\delta}}^{(i)}, \hat{\boldsymbol{\epsilon}}^{(i)}\}_{i=1}^{N}), \rho(\{\hat{\boldsymbol{\delta}}^{(i,1)}, \hat{\boldsymbol{\epsilon}}^{(i,1)}\}_{i=1}^{N}), \ldots, \rho(\{\hat{\boldsymbol{\delta}}^{(i,M)}, \hat{\boldsymbol{\epsilon}}^{(i,M)}\}_{i=1}^{N})$$

where $(\hat{\boldsymbol{\delta}}^{(i)}, \hat{\boldsymbol{\epsilon}}^{(i)})$ is the ith sample of $\mathcal{D}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}$ and $(\hat{\boldsymbol{\delta}}^{(i,m)}, \hat{\boldsymbol{\epsilon}}^{(i,m)})$ is the ith sample of dataset $\mathcal{D}^{(m)}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}$. As ρ is deterministic, it suffices to show that the following sequence is exchangeable conditional on $\{(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^{N}$:

$$\{\hat{\boldsymbol{\delta}}^{(i)}, \hat{\boldsymbol{\epsilon}}^{(i)}\}_{i=1}^{N}, \{\hat{\boldsymbol{\delta}}^{(i,1)}, \hat{\boldsymbol{\epsilon}}^{(i,1)}\}_{i=1}^{N}, \ldots, \{\hat{\boldsymbol{\delta}}^{(i,M)}, \hat{\boldsymbol{\epsilon}}^{(i,M)}\}_{i=1}^{N}$$

Note that $\hat{\boldsymbol{\delta}}^{(i)}, \hat{\boldsymbol{\epsilon}}^{(i)} \sim \hat{q}(\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\delta}} \mid \mathbf{x}^{(i)}, \mathbf{y}^{(i)}, \mathbf{z}^{(i)})$. This means that the estimated information residuals can be written as $\hat{\boldsymbol{\delta}}^{(i)}, \hat{\boldsymbol{\epsilon}}^{(i)} = h(\boldsymbol{\alpha}^{(i)}, \mathbf{x}^{(i)}, \mathbf{y}^{(i)}, \mathbf{z}^{(i)}; \theta_{\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}})$, where h is a deterministic function (see appendix A of Trivedi and Zimmer (2007)), and $\theta_{\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}}$ denotes the fact that $\hat{q}$ is trained on the dataset $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$. Rewriting the sampling process as a function of independent noise is similar in spirit to the reparameterization trick used in variational inference (Kingma et al., 2015).

In this alternative representation, $\boldsymbol{\alpha}^{(i)}$ is a sample of exogenous variable $\boldsymbol{\alpha}$ that represents the noise in $\hat{q}$. Using the same notation, $\hat{\boldsymbol{\delta}}^{(i,m)}, \hat{\boldsymbol{\epsilon}}^{(i,m)} = h(\boldsymbol{\alpha}^{(i,m)}, \widetilde{\mathbf{x}}^{(i,m)}, \mathbf{y}^{(i)}, \mathbf{z}^{(i)}; \theta_{\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}})$, where $\widetilde{\mathbf{x}}^{(i,m)}$ is the ith sample of the mth null dataset $\widetilde{\mathbf{X}}^{(m)}$ and $\boldsymbol{\alpha}^{(i,m)}$ is another independent sample of $\boldsymbol{\alpha}$. This means the above sequence can be written as:

$$\begin{aligned}
&\{h(\boldsymbol{\alpha}^{(i)}, \mathbf{x}^{(i)}, \mathbf{y}^{(i)}, \mathbf{z}^{(i)}; \theta_{\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}})\}_{i=1}^{N}, \\
&\{h(\boldsymbol{\alpha}^{(i,1)}, \widetilde{\mathbf{x}}^{(i,1)}, \mathbf{y}^{(i)}, \mathbf{z}^{(i)}; \theta_{\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}})\}_{i=1}^{N}, \\
&\vdots \\
&\{h(\boldsymbol{\alpha}^{(i,M)}, \widetilde{\mathbf{x}}^{(i,M)}, \mathbf{y}^{(i)}, \mathbf{z}^{(i)}; \theta_{\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}})\}_{i=1}^{N}.
\end{aligned}$$

Since h is deterministic and learned from $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$, exchangeability of the set of random variables above reduces to exchangeability of the following:

$$\begin{aligned}
&(\{\boldsymbol{\alpha}^{(i)}, \mathbf{x}^{(i)}, \mathbf{y}^{(i)}, \mathbf{z}^{(i)}\}_{i=1}^{N}, \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}), \\
&(\{\boldsymbol{\alpha}^{(i,1)}, \widetilde{\mathbf{x}}^{(i,1)}, \mathbf{y}^{(i)}, \mathbf{z}^{(i)}\}_{i=1}^{N}, \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}), \\
&\vdots \\
&(\{\boldsymbol{\alpha}^{(i,M)}, \widetilde{\mathbf{x}}^{(i,M)}, \mathbf{y}^{(i)}, \mathbf{z}^{(i)}\}_{i=1}^{N}, \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}})
\end{aligned} \tag{4}$$

Now, recall that $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}} = \{\widetilde{\mathbf{x}}^{(i)}, \mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^{N}$ where each $\widetilde{\mathbf{x}}^{(i)}$ is a random sample from $p(\mathbf{x} \mid \mathbf{z} = \mathbf{z}^{(i)})$ and note that the only dependence between $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$ and $\mathbf{x}^{(i)}, \mathbf{x}^{(i,m)}$ is through $\mathbf{y}^{(i)}, \mathbf{z}^{(i)}$. Then, collecting the identically distributed samples within in each element in the sequence in eq. (4)

$$\begin{aligned}
\mathbf{x}^{(i)} \perp\!\!\!\perp \widetilde{\mathbf{x}}^{(i)} \mid \mathbf{y}_i, \mathbf{z}_i \implies \{\mathbf{x}^{(i)}\}_{i=1}^{N} \perp\!\!\!\perp \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}} \mid \{(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^{N}. \\
\forall m \quad \widetilde{\mathbf{x}}^{(i,m)} \perp\!\!\!\perp \widetilde{\mathbf{x}}^{(i)} \mid \mathbf{y}_i, \mathbf{z}_i \implies \{\widetilde{\mathbf{x}}^{(i,m)}\}_{i=1}^{N} \perp\!\!\!\perp \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}} \mid \{(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^{N}.
\end{aligned}$$

This fact imply these two equalities in distribution (between the conditioning set containing $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$ and otherwise):

$$\begin{aligned}
\{\mathbf{x}^{(i)}\}_{i=1}^{N} \mid \{(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^{N}, \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}} &=^{d} \{\mathbf{x}^{(i)}\}_{i=1}^{N} \mid \{(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^{N}, \\
\forall m \quad \{\mathbf{x}^{(i,m)}\}_{i=1}^{N} \mid \{(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^{N}, \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}} &=^{d} \{\mathbf{x}^{(i,m)}\}_{i=1}^{N} \mid \{(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^{N}
\end{aligned} \tag{5}$$

Now, the two RHS's above are equal in distribution under the null hypothesis: under $\mathcal{H}_0$, $p(\mathbf{x} \mid \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z}, \mathbf{y})$, which means that the distribution of $\mathbf{x}^{(i)}$ is equal to the distribution of $\widetilde{\mathbf{x}}^{(i,m)}$ given $\{(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^N$. This fact means the LHS's in eq. (5) are equal which implies the following equality in distribution

$$\{\mathbf{x}^{(i)}\}_{i=1}^N \mid \{(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^N, \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}} \quad =^d \quad \{\mathbf{x}^{(i,m)}\}_{i=1}^N \mid \{(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^N, \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}} \tag{6}$$

Then, recalling that $\boldsymbol{\alpha}^{(i)}$ and $\boldsymbol{\alpha}^{(i,m)}$ for all m are exogenous random variables, eq. (6) implies that given $\{(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^N, \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$, the random variable $\{(\boldsymbol{\alpha}^{(i)}, \mathbf{x}^{(i)})\}_{i=1}^N$ is distributed identically to $\{\boldsymbol{\alpha}^{(i,m)}, \widetilde{\mathbf{x}}^{(i,m)}\}_{i=1}^N$ for any m. Finally, as $\{(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^N, \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$ is constant across each element of the sequence in eq. (4), the sequence is exchangeable.

□

Corollary 2.1. *Let $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ be drawn from any distribution $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ and $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ consist of N iid samples from this distribution. If $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$, then for any measure of marginal dependence $\rho : (\mathbb{R} \times \mathbb{R})^N \to \mathbb{R}$, let $\hat{\mathbf{p}}$ the p-value computed using any residuals $\hat{\boldsymbol{\epsilon}} = u(\mathbf{x}, \mathbf{z})$ and $\hat{\boldsymbol{\delta}} = v(\mathbf{y}, \mathbf{z})$ where u, v are learned from the null-dataset $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}} \sim p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z}, \mathbf{y})$. Then, $\hat{\mathbf{p}}$ will stochastically dominate a Uniform$(0, 1)$ random variable.*

Proof. The one property of DIET used in proving prop. 1 is that $\hat{q}$ is learned using the dataset $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$. This gives us the property that under the null hypothesis

$$\{\mathbf{x}^{(i)}\}_i^N \perp\!\!\!\perp \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}} \mid \{\mathbf{y}^{(i)}, \mathbf{z}^{(i)}\}_i^N\} \qquad \{\widetilde{\mathbf{x}}^{(i,m)}\}_i^N \perp\!\!\!\perp \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}} \mid \{\mathbf{y}^{(i)}, \mathbf{z}^{(i)}\}_i^N\},$$

because 1) the dependence between variables $\mathbf{x}^{(i)}$ or $\widetilde{\mathbf{x}}^{(i,m)}$ and $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}} = \{\widetilde{\mathbf{x}}^{(i)}, \mathbf{y}^{(i)}, \mathbf{z}^{(i)}\}_{i=1}^N$ and is only due to $\{\mathbf{y}^{(i)}, \mathbf{z}^{(i)}\}_i^N$, which we condition on, 2) the $\widetilde{\mathbf{x}}^{(i)}$ samples in $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$ are independent samples drawn from $p(\mathbf{x} \mid \mathbf{z} = \mathbf{z}^{(i)})$. These two properties imply the following independencies:

$$\mathbf{x}^{(i)} \perp\!\!\!\perp \widetilde{\mathbf{x}}^{(i)} \mid \mathbf{y}_i, \mathbf{z}_i \implies \{\mathbf{x}^{(i)}\}_{i=1}^N \perp\!\!\!\perp \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}} \mid \{(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^N.$$
$$\forall m \quad \mathbf{x}^{(i,m)} \perp\!\!\!\perp \widetilde{\mathbf{x}}^{(i)} \mid \mathbf{y}_i, \mathbf{z}_i \implies \{\widetilde{\mathbf{x}}^{(i,m)}\}_{i=1}^N \perp\!\!\!\perp \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}} \mid \{(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^N.$$

For any u, v learned from $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}} \sim p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z}, \mathbf{y})$, the same properties hold because samples from $\hat{q}(\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\delta}} \mid \mathbf{y}, \mathbf{x}, \mathbf{z})$ are produced as $\hat{\boldsymbol{\epsilon}}^{(i)} = u(\mathbf{x}^{(i)}, \mathbf{z}^{(i)})$ and $\hat{\boldsymbol{\delta}}^{(i)} = v(\mathbf{y}^{(i)}, \mathbf{z}^{(i)})$ and u, v are learned from $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$.

Thus, theorem 1 will hold for any functions u, v learned from $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$ or using data from $p(\mathbf{x}, \mathbf{z})$ and $p(\mathbf{y}, \mathbf{z})$ respectively; then, using $\hat{\boldsymbol{\epsilon}} = u(\mathbf{x}, \mathbf{z})$ and $\hat{\boldsymbol{\delta}} = v(\mathbf{y}, \mathbf{z})$ to compute a p-value using

$$\hat{\mathbf{p}} = \frac{1}{M+1}\left(1 + \sum_{m=1}^{M} \mathbb{1}\left[\rho(\mathcal{D}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}) \geq \rho(\mathcal{D}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}^{(m)})\right]\right)$$

will result in a super-uniform $\hat{\mathbf{p}}$. □

A.3 Proof of theorem 1

Theorem 1. *Let $F_{\cdot|\mathbf{z}}(\cdot \mid \mathbf{z})$ denote the conditional* CDF *for the distribution $p(\cdot \mid \mathbf{z})$. Let $\boldsymbol{\epsilon} = F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z})$ and $\boldsymbol{\delta} = F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z})$ be random variables defined over $(\mathbf{x}, \mathbf{z})$ and $(\mathbf{y}, \mathbf{z})$ respectively. Assume F is invertible in the first argument and $(\boldsymbol{\epsilon}, \boldsymbol{\delta}) \perp\!\!\!\perp \mathbf{z}$. If there exists a marginal independence test $\psi : (\mathbb{R} \times \mathbb{R})^N \times [0, 1] \to \{0, 1\}$ that uses a measure of dependence ρ and achieves power greater than $\alpha \in [0, 1]$, then* DIET *equipped with ρ and the conditional* CDFs *$F(\cdot \mid \mathbf{z})$ is a conditional independence test with power greater than α for data drawn from $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$.*

Proof. To test the conditional independence relationship $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$, DIET tests the marginal independence between $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$. The aim of this proof is to show that $\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\delta}$ if and only if $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$. If this reduction holds, then under the alternate hypothesis $\mathcal{H}_1$ where $\mathbf{x} \not\perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$, the distribution of the test statistic $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}})$ will be different from the distribution of each of the null statistics $T(\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)})$. Then, given any marginal independence test that achieves power $> \alpha$ with statistic ρ, DIET with the same statistic is a *conditional* independence test with power $> \alpha$.

The proof is structured in the following manner. First, we will show that using the null data $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)}$, the sampled values of $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ will be independent. Then, we will show that using the true data $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$, the sampled values of $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ will be dependent. Finally, we discuss how the existence of a marginal independence test with power $> \alpha$ implies that DIET will also achieve power $> \alpha$ using data $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$.

Prerequisites. We first outline some properties will be used in both the null statistics and the test statistic section.

$$\begin{aligned} p(\boldsymbol{\epsilon}, \mathbf{z}) &= \int p(\boldsymbol{\epsilon}, \boldsymbol{\delta}, \mathbf{z}) d\boldsymbol{\delta} && \text{by marginalization} \\ &= \int p(\boldsymbol{\epsilon}, \boldsymbol{\delta}) p(\mathbf{z}) d\boldsymbol{\delta} && \text{by data distribution} \\ &= p(\boldsymbol{\epsilon}) p(\mathbf{z}) \end{aligned}$$

Thus,

$$\boldsymbol{\epsilon} \perp\!\!\!\perp \mathbf{z} \tag{7}$$
$$\boldsymbol{\delta} \perp\!\!\!\perp \mathbf{z}. \tag{8}$$

Null statistics $T(\mathcal{D}_{\tilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)})$. Recall that in each of the null datasets, the following factorization of the data distribution $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ holds by construction:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z}). \tag{9}$$

We can use this property to make the following sequence of deductions. Letting $p(\boldsymbol{\epsilon}, \boldsymbol{\delta}, \mathbf{z})$ be the distribution implied by $(\boldsymbol{\epsilon}, \boldsymbol{\delta}, \mathbf{z})$,

$$\begin{aligned} p(\boldsymbol{\epsilon}, \boldsymbol{\delta}, \mathbf{z}) &= \int p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) p(\mathbf{x}, \mathbf{y}, \mathbf{z}) d\mathbf{x} d\mathbf{y} \\ &= \int p(\boldsymbol{\epsilon} \mid \mathbf{x}, \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{y}, \mathbf{z}) p(\mathbf{x}, \mathbf{y}, \mathbf{z}) d\mathbf{x} d\mathbf{y} && \boldsymbol{\epsilon} \text{ and } \boldsymbol{\delta} \text{ are each functions of } \mathbf{z} \text{ and either } \mathbf{x} \text{ or } \mathbf{y} \\ &= \int p(\boldsymbol{\epsilon} \mid \mathbf{x}, \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{y}, \mathbf{z}) p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z}) d\mathbf{x} d\mathbf{y} && \text{by eq. (9)} \\ &= \int p(\boldsymbol{\epsilon}, \mathbf{x} \mid \mathbf{z}) p(\boldsymbol{\delta}, \mathbf{y} \mid \mathbf{z}) p(\mathbf{z}) d\mathbf{x} d\mathbf{y} \\ &= p(\boldsymbol{\epsilon} \mid \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{z}) p(\mathbf{z}) \\ p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{z}) &= p(\boldsymbol{\epsilon} \mid \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{z}). \end{aligned}$$

The distribution of $(\mathbf{y}, \mathbf{z})$ under the null is the same as distribution of $(\mathbf{y}, \mathbf{z})$ in the data. Then since $\boldsymbol{\delta} \perp\!\!\!\perp \mathbf{z}$ (eq. 8) holds in the data distribution, the independence of $\boldsymbol{\delta}$ and $\mathbf{z}$ also holds under the null distribution eq. (9):

$$\boldsymbol{\delta} \perp\!\!\!\perp \mathbf{z} \qquad \text{where } (\mathbf{x}, \mathbf{y}, \mathbf{z}) \sim p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z}); \quad \boldsymbol{\delta} = F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z}).$$

Using the same logic, eq. (7) implies

$$\boldsymbol{\epsilon} \perp\!\!\!\perp \mathbf{z} \qquad \text{where } (\mathbf{x}, \mathbf{y}, \mathbf{z}) \sim p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z}); \quad \boldsymbol{\epsilon} = F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z}).$$

Using the above facts,

$$\begin{aligned} p(\boldsymbol{\epsilon}, \boldsymbol{\delta}) &= \int p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{z}) p(\mathbf{z}) d\mathbf{z} && \text{by marginalization} \\ &= \int p(\boldsymbol{\epsilon} \mid \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{z}) p(\mathbf{z}) d\mathbf{z} \\ &= \int p(\boldsymbol{\epsilon}) p(\boldsymbol{\delta}) p(\mathbf{z}) d\mathbf{z} \\ &= p(\boldsymbol{\epsilon}) p(\boldsymbol{\delta}). \end{aligned}$$

Therefore, when using a null dataset $\mathcal{D}_{\tilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)}$, $\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\delta}$.

Test statistic $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}})$**.** Under $\mathcal{H}_1$, $\mathbf{x} \not\perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$. In such cases, the sampled values of $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ using $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ must be dependent. Specifically, the following sequence of implications must hold:

$$\mathbf{x} \not\perp\!\!\!\perp \mathbf{y} \mid \mathbf{z} \Rightarrow \boldsymbol{\epsilon} \not\perp\!\!\!\perp \boldsymbol{\delta} \mid \mathbf{z} \Rightarrow \boldsymbol{\epsilon} \not\perp\!\!\!\perp \boldsymbol{\delta}.$$

The first implication follows because both $F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z})$ and $F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z})$ are invertible for any fixed value of $\mathbf{z}$. Next we prove the second implication. This is equivalent to:

$$\boldsymbol{\delta} \perp\!\!\!\perp \boldsymbol{\epsilon} \Rightarrow \boldsymbol{\delta} \perp\!\!\!\perp \boldsymbol{\epsilon} \mid \mathbf{z}.$$

We know that $p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{z}) = p(\boldsymbol{\epsilon}, \boldsymbol{\delta})$. It follows that $\boldsymbol{\delta} \perp\!\!\!\perp \boldsymbol{\epsilon} \Rightarrow \boldsymbol{\delta} \perp\!\!\!\perp \boldsymbol{\epsilon} \mid \mathbf{z}$:

$$\begin{aligned} p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{z}) &= p(\boldsymbol{\epsilon}, \boldsymbol{\delta}) \\ &= p(\boldsymbol{\epsilon})p(\boldsymbol{\delta}) && \text{since } \boldsymbol{\delta} \perp\!\!\!\perp \boldsymbol{\epsilon} \\ &= p(\boldsymbol{\epsilon} \mid \mathbf{z})p(\boldsymbol{\delta} \mid \mathbf{z}) && \text{by eqs. (7) and (8)} \end{aligned}$$

We have thus far established that under $\mathcal{H}_1$, $\boldsymbol{\delta} \not\perp\!\!\!\perp \boldsymbol{\epsilon}$, but under $\mathcal{H}_0$, $\boldsymbol{\delta} \perp\!\!\!\perp \boldsymbol{\epsilon}$. Now, consider $\psi(\mathcal{D}_{\boldsymbol{\epsilon},\boldsymbol{\delta}}, \alpha) : (\mathbb{R} \times \mathbb{R})^N \times [0,1] \to \{0,1\}$, a marginal independence test that uses statistic $\rho : (\mathbb{R} \times \mathbb{R})^N \to \mathbb{R}$ and has power greater than level α. This means that there exists a rejection region $R_\alpha = \{\mathcal{D} \in (\mathbb{R} \times \mathbb{R})^N : \psi(\mathcal{D}, \alpha) = 1\}$ where $\mathbb{P}_{\mathcal{H}_1}(R_\alpha) \geq \mathbb{P}_{\mathcal{H}_0}(R_\alpha)$. In other words, for a sample size of N and statistic ρ there is sufficient evidence to reject the null hypothesis.

Then, DIET equipped with ρ, $F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z})$, and $F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z})$ is a conditional independence test $\zeta(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}, \alpha) : (\mathbb{R} \times \mathbb{R} \times \mathbb{R}^{d_\mathbf{z}})^N \times [0,1] \to \{0,1\}$ with rejection region $S_\alpha = \{\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}} \in (\mathbb{R} \times \mathbb{R} \times \mathbb{R}^p)^N : \zeta(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}, \alpha) = 1\}$ such that $\mathbb{P}_{\mathcal{H}_1}(S_\alpha) \geq \mathbb{P}_{\mathcal{H}_0}(S_\alpha)$. This follows directly from the previous fact because DIET uses the marginal dependence $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ to test the conditional independence between $\mathbf{x}$ and $\mathbf{y}$ given $\mathbf{z}$.

Thus, if there is a marginal test that achieves power greater than α, then DIET under the conditions of theorem 1 will also achieve power greater than α.

□

A.3.1 Example family of data generating proceses that satisfy the core assumption in theorem 1

Here, we specify a family of data generating processes that satisfies $(\boldsymbol{\epsilon}, \boldsymbol{\delta}) \perp\!\!\!\perp \mathbf{z}$. Let $\mathbf{e}, \mathbf{d}$ be any continuously distributed random variables with contiguous support and let the joint distribution over $\mathbf{e}, \mathbf{d}, \mathbf{z}$ be

$$p(\mathbf{e}, \mathbf{d}, \mathbf{z}) = p(\mathbf{e}, \mathbf{d})p(\mathbf{z}).$$

For any pair of functions $f(\cdot,\cdot), g(\cdot,\cdot)$ that are continuous and strictly monotonic in the first argument, let samples from $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ be generated as:

$$\mathbf{z} \sim p(\mathbf{z}),\ \mathbf{e}, \mathbf{d} \sim p(\mathbf{e}, \mathbf{d}),\ \mathbf{x} = f(\mathbf{e}, \mathbf{z}),\ \mathbf{y} = g(\mathbf{d}, \mathbf{z}).$$

The core assumption holds on all $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ with the above form. First, we express $\boldsymbol{\epsilon}$ as a deterministic function of $\mathbf{e}$: almost surely under $p(\mathbf{y}, \mathbf{x}, \mathbf{z})$: with $f_z^{-1}(x, z)$ as the inverse of the function f in the first argument with z fixed

$$\boldsymbol{\epsilon} = F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z}) = F_{f_z^{-1}(\mathbf{x},\mathbf{z})|\mathbf{z}}(f_z^{-1}(\mathbf{x}, \mathbf{z}) \mid \mathbf{z}) = F_{\mathbf{e}|\mathbf{z}}(\mathbf{e} \mid \mathbf{z}) = F_\mathbf{e}(\mathbf{e}).$$

The second equality holds as CDFs are invariant to strictly monotonic transformations of the underlying random variables and the fourth due to the independence $\mathbf{e} \perp\!\!\!\perp \mathbf{z}$. Similarly $\boldsymbol{\delta} = F_\mathbf{d}(\mathbf{d})$. In turn, $\boldsymbol{\delta}, \boldsymbol{\epsilon}$ are deterministic functions of $\mathbf{d}, \mathbf{e}$ respectively, and the core assumption holds:

$$(\mathbf{d}, \mathbf{e}) \perp\!\!\!\perp \mathbf{z} \implies (F_\mathbf{e}(\mathbf{e}), F_\mathbf{d}(\mathbf{d})) \perp\!\!\!\perp \mathbf{z} \implies (\boldsymbol{\epsilon}, \boldsymbol{\delta}) \perp\!\!\!\perp \mathbf{z}.$$

A.4 Generalizing assumptions for distillation

In this section we consider data generating processes of the following form:

$$\mathbf{z} \sim p(\mathbf{z}) \qquad (\mathbf{e}, \mathbf{d}) \sim p(\mathbf{e}, \mathbf{d}) \qquad \mathbf{x} = f(\mathbf{e}, \mathbf{z}) \qquad \mathbf{y} = g(\mathbf{d}, \mathbf{z}).$$

The goal of a distillation procedure like DIET or the d_0-CRT is to first estimate $\mathbf{e}$ and $\mathbf{d}$ from samples of $(\mathbf{x}, \mathbf{y}, \mathbf{z})$, then test the marginal independence of these estimates: $\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\delta}$. Since $\mathbf{e}$ and $\mathbf{d}$ are unobserved, samples in $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ map to a distribution $p(\mathbf{d}, \mathbf{e} \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$ over the possible values of $(\mathbf{e}, \mathbf{d})$. The distribution $p(\mathbf{e}, \mathbf{d} \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$ is also unknown; it must be estimated using an estimator $\hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$.

However, not all $\boldsymbol{\epsilon}, \boldsymbol{\delta} \sim \hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$ will yield power to reject the null hypothesis $\mathcal{H}_0 : \mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$. In some cases $\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\delta}$ but $\mathbf{x} \not\perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$. Consider this example from Puli and Ranganath (2020). Let $\mathbf{x} = \boldsymbol{\epsilon}$ and $\mathbf{y} = \boldsymbol{\delta}$, let $\boldsymbol{\epsilon} \sim \text{Uniform}(0, 1)$, $\boldsymbol{\delta} \sim \text{Uniform}(0, 1)$, and

$$\mathbf{z} = \begin{cases} \boldsymbol{\epsilon} + \boldsymbol{\delta} & \text{if } \boldsymbol{\epsilon} + \boldsymbol{\delta} \leq 1 \\ \boldsymbol{\epsilon} + \boldsymbol{\delta} - 1 & \text{otherwise} \end{cases}.$$

In this example, $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are independent of each other, but $\mathbf{x}$ and $\mathbf{y}$ are clearly dependent given $\mathbf{z}$. The following theorem, theorem 2, gives sufficient conditions on $\hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$ to ensure that $\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\delta}$ if and only if $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$. We later show that the only way to satisfy the conditions in theorem 2 are through assumptions on the data generating process.

Theorem 2. *Consider a data generating process of the following form:*

$$\mathbf{z} \sim p(\mathbf{z}), \quad \mathbf{e}, \mathbf{d} \sim p(\mathbf{e}, \mathbf{d}), \quad \mathbf{x} = f(\mathbf{e}, \mathbf{z}) \quad \mathbf{y} = g(\mathbf{d}, \mathbf{z}).$$

Let $\boldsymbol{\epsilon}, \boldsymbol{\delta}$ be distributed according to:

$$(\boldsymbol{\epsilon}, \boldsymbol{\delta}, \mathbf{x}, \mathbf{y}, \mathbf{z}) \sim \hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) p(\mathbf{x}, \mathbf{y}, \mathbf{z}).$$

Further let,

$$\hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\boldsymbol{\epsilon} \mid \mathbf{x}, \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{y}, \mathbf{z}), \qquad \text{(factorization)}$$

$$\exists \tilde{f}, \tilde{g} \qquad \mathbf{x} \overset{a.s.}{=} \tilde{f}(\boldsymbol{\epsilon}, \mathbf{z}), \quad \mathbf{y} \overset{a.s.}{=} \tilde{g}(\boldsymbol{\delta}, \mathbf{z}), \qquad \text{(reconstruction)}$$

$$(\mathbf{d}, \boldsymbol{\delta}) \perp\!\!\!\perp \mathbf{z} \qquad (\mathbf{e}, \boldsymbol{\epsilon}) \perp\!\!\!\perp \mathbf{z}. \qquad \text{(joint independence)}$$

Let $\psi(\mathcal{D}_{\boldsymbol{\epsilon},\boldsymbol{\delta}}, \alpha) : (\mathbb{R} \times \mathbb{R})^N \times [0, 1] \to \{0, 1\}$ be a marginal independence test that uses statistic $\rho : (\mathbb{R} \times \mathbb{R})^N \to \mathbb{R}$ and has power greater than α. Let $\mathcal{D}_{\boldsymbol{\epsilon},\boldsymbol{\delta}}$ be a dataset of N samples of $(\boldsymbol{\epsilon}, \boldsymbol{\delta})$ generated using $\hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$ and $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$. Then, ψ using $\mathcal{D}_{\boldsymbol{\epsilon},\boldsymbol{\delta}}$ and ρ is also a conditional test of independence for $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$ with power greater than α.

Proof. The core of this proof is to show that if factorization, reconstruction, and joint independence are satisfied, then

$$\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z} \Leftrightarrow \boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\delta}.$$

If this reduction is possible, then under $\mathcal{H}_1$, $\boldsymbol{\epsilon} \not\perp\!\!\!\perp \boldsymbol{\delta}$, but under $\mathcal{H}_0$, $\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\delta}$. This implies that the distribution of the marginal dependence test statistic $\rho(\mathcal{D}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}})$ is different from that of each null statistic $\rho(\mathcal{D}^{(m)}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}})$. Thus, the p-value computed by ψ will be close to 0:

$$\hat{\mathbf{p}} = \frac{1}{M+1}\left(1 + \sum_{m=1}^{M} \mathbb{1}\left(\rho(\mathcal{D}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}}) \leq \rho(\mathcal{D}^{(m)}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}})\right)\right).$$

Let $p(\mathbf{d}, \mathbf{e}, \boldsymbol{\epsilon}, \boldsymbol{\delta}, \mathbf{z})$ be a distribution over variables $\mathbf{d}, \mathbf{e}, \boldsymbol{\epsilon}, \boldsymbol{\delta}, \mathbf{z}$. The variables $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ are samples from $\hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$. For simplicity, we show the proof of theorem 2 when all random variables are continuous, but the same reasoning holds for discrete random variables.

Null statistics. For null statistics $\rho(\mathcal{D}^{(m)}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}})$ computed using null data ${\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}}^{(m)}$, $\boldsymbol{\delta}, \boldsymbol{\epsilon} \sim \hat{q}(\boldsymbol{\delta}, \boldsymbol{\epsilon} \mid \widetilde{\mathbf{x}}, \mathbf{y}, \mathbf{z})$ must be independent. In the null data, $\widetilde{\mathbf{x}} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$ by construction, so the following must hold:

$$\widetilde{\mathbf{x}} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z} \Rightarrow \boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\delta}. \tag{10}$$

We show this fact by manipulating the distribution $\hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \widetilde{\mathbf{x}}, \mathbf{y}, \mathbf{z}) p(\widetilde{\mathbf{x}}, \mathbf{y}, \mathbf{z})$. In this proof, we write $\hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \widetilde{\mathbf{x}}, \mathbf{y}, \mathbf{z})$ as $p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \widetilde{\mathbf{x}}, \mathbf{y}, \mathbf{z})$ to simplify the notation:

$$p(\boldsymbol{\epsilon}, \boldsymbol{\delta}, \mathbf{z}) = \int p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \widetilde{\mathbf{x}}, \mathbf{y}, \mathbf{z}) p(\widetilde{\mathbf{x}}, \mathbf{y}, \mathbf{z}) d\widetilde{\mathbf{x}} d\mathbf{y}$$

$$
\begin{aligned}
&= \int p(\boldsymbol{\epsilon} \mid \widetilde{\mathbf{x}}, \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{y}, \mathbf{z}) p(\widetilde{\mathbf{x}}, \mathbf{y}, \mathbf{z}) d\widetilde{\mathbf{x}} d\mathbf{y} && \text{By factorization} \\
&= \int p(\boldsymbol{\epsilon} \mid \mathbf{x}, \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{y}, \mathbf{z}) p(\widetilde{\mathbf{x}} \mid \mathbf{z}) p(\mathbf{y} \mid \mathbf{z}) p(\mathbf{z}) d\widetilde{\mathbf{x}} d\mathbf{y} && \text{In } \mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}\ \widetilde{\mathbf{x}} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z} \\
&= \int p(\boldsymbol{\epsilon}, \widetilde{\mathbf{x}} \mid \mathbf{z}) p(\boldsymbol{\delta}, \mathbf{y} \mid \mathbf{z}) p(\mathbf{z}) d\widetilde{\mathbf{x}} d\mathbf{y} \\
&= p(\boldsymbol{\epsilon} \mid \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{z}) p(\mathbf{z})
\end{aligned}
$$

Consequently, $p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{z}) = p(\boldsymbol{\delta} \mid \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{z}) \Leftrightarrow \boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\delta} \mid \mathbf{z}$. Here, if $\boldsymbol{\delta} \perp\!\!\!\perp \mathbf{z}$ and $\boldsymbol{\epsilon} \perp\!\!\!\perp \mathbf{z}$, then

$$
\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\delta} \mid \mathbf{z} \Rightarrow \boldsymbol{\epsilon} \perp\!\!\!\perp (\boldsymbol{\delta}, \mathbf{z}) \Rightarrow \boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\delta}
$$

Thus, if $\widetilde{\mathbf{x}} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$, as is the case in the computation of each of the null statistics $\rho(\mathcal{D}^{(m)}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}})$, then for $\boldsymbol{\delta}, \boldsymbol{\epsilon} \sim \hat{q}(\boldsymbol{\delta}, \boldsymbol{\epsilon} \mid \widetilde{\mathbf{x}}, \mathbf{y}, \mathbf{z})$, $\boldsymbol{\delta} \perp\!\!\!\perp \boldsymbol{\epsilon}$.

Test statistic under $\mathcal{H}_1$. For the test statistic $\rho(\mathcal{D}_{\hat{\boldsymbol{\epsilon}},\hat{\boldsymbol{\delta}}})$ computed using the data $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$, $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ must be dependent. Under $\mathcal{H}_1$, $\mathbf{x} \not\perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$, so the following sequence of implications must hold:

$$
\mathbf{x} \not\perp\!\!\!\perp \mathbf{y} \mid \mathbf{z} \Rightarrow \tilde{f}(\boldsymbol{\epsilon}, \mathbf{z}) \not\perp\!\!\!\perp \tilde{g}(\boldsymbol{\delta}, \mathbf{z}) \mid \mathbf{z} \Rightarrow \boldsymbol{\epsilon} \not\perp\!\!\!\perp \boldsymbol{\delta} \mid \mathbf{z} \Rightarrow \boldsymbol{\epsilon} \not\perp\!\!\!\perp \boldsymbol{\delta}. \tag{11}
$$

The first implication follows directly from reconstruction, the second holds because $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are the only sources of variance when $\mathbf{z}$ is fixed. This last implication is equivalent to the following statement, which we will subsequently prove:

$$
\boldsymbol{\delta} \perp\!\!\!\perp \boldsymbol{\epsilon} \Rightarrow \boldsymbol{\delta} \perp\!\!\!\perp \boldsymbol{\epsilon} \mid \mathbf{z}.
$$

First, note the following properties. Using joint independence, we show that the distribution $p(\boldsymbol{\delta}, \mathbf{z})$ factorizes, implying that $\boldsymbol{\delta}$ and $\mathbf{z}$ are marginally independent:

$$
p(\boldsymbol{\delta}, \mathbf{z}) = \int p(\mathbf{d}, \boldsymbol{\delta}, \mathbf{z}) d\mathbf{d} = \int p(\mathbf{d}, \boldsymbol{\delta}) p(\mathbf{z}) d\mathbf{d} = p(\boldsymbol{\delta}) p(\mathbf{z}), \tag{12}
$$

$$
p(\boldsymbol{\epsilon}, \mathbf{z}) = \int p(\mathbf{e}, \boldsymbol{\epsilon}, \mathbf{z}) d\mathbf{d} = \int p(\mathbf{e}, \boldsymbol{\epsilon}) p(\mathbf{z}) d\mathbf{e} = p(\boldsymbol{\epsilon}) p(\mathbf{z}). \tag{13}
$$

Further, joint independence implies the following:

$$
\frac{p(\boldsymbol{\delta}, \mathbf{d}, \mathbf{z})}{p(\mathbf{d})} = \frac{p(\boldsymbol{\delta}, \mathbf{d}) p(\mathbf{z})}{p(\mathbf{d})} = p(\boldsymbol{\delta} \mid \mathbf{d}) p(\mathbf{z}) = p(\boldsymbol{\delta} \mid \mathbf{d}) p(\mathbf{z} \mid \mathbf{d}) \qquad \text{Since } \mathbf{z} \perp\!\!\!\perp \mathbf{d} \text{ by definition} \tag{14}
$$

$$
\frac{p(\boldsymbol{\epsilon}, \mathbf{e}, \mathbf{z})}{p(\mathbf{e})} = \frac{p(\boldsymbol{\epsilon}, \mathbf{e}) p(\mathbf{z})}{p(\mathbf{e})} = p(\boldsymbol{\epsilon} \mid \mathbf{e}) p(\mathbf{z}) = p(\boldsymbol{\epsilon} \mid \mathbf{e}) p(\mathbf{z} \mid \mathbf{e}) \qquad \text{Since } \mathbf{z} \perp\!\!\!\perp \mathbf{e} \text{ by definition} \tag{15}
$$

Next, note that factorization implies:

$$
\begin{aligned}
p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{z}, \mathbf{e}, \mathbf{d}) &= p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{e}, \mathbf{d}) && \mathbf{x} \text{ and } \mathbf{y} \text{ are fully determined by } (\mathbf{z}, \mathbf{d}, \mathbf{e}) \\
&= p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) && (\boldsymbol{\epsilon}, \boldsymbol{\delta}) \text{ are functions of only } (\mathbf{x}, \mathbf{y}, \mathbf{z}) \text{ and exogenous noise} \\
&= p(\boldsymbol{\epsilon} \mid \mathbf{x}, \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{y}, \mathbf{z}) && \text{factorization assumption} \\
&= p(\boldsymbol{\epsilon} \mid \mathbf{x}, \mathbf{z}, \mathbf{e}) p(\boldsymbol{\delta} \mid \mathbf{y}, \mathbf{z}, \mathbf{d}) && \boldsymbol{\epsilon}, \boldsymbol{\delta} \text{ are functions of } (\mathbf{x}, \mathbf{z}) \text{ and } (\mathbf{y}, \mathbf{z}) \text{ respectively and exogenous noise} \\
&= p(\boldsymbol{\epsilon} \mid \mathbf{e}, \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{d}, \mathbf{z}) && \mathbf{x} = f(\mathbf{e}, \mathbf{z}), \mathbf{y} = g(\mathbf{d}, \mathbf{z}).
\end{aligned} \tag{16}
$$

Using the above facts, we then show that $p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{z}) = p(\boldsymbol{\epsilon}, \boldsymbol{\delta})$:

$$
\begin{aligned}
p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{z}) &= \int p(\boldsymbol{\epsilon}, \boldsymbol{\delta}, \mathbf{e}, \mathbf{d} \mid \mathbf{z}) d\mathbf{e} d\mathbf{d} && \text{By marginalization} \\
&= \int p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{z}, \mathbf{e}, \mathbf{d}) p(\mathbf{e}, \mathbf{d} \mid \mathbf{z}) d\mathbf{e} d\mathbf{d} \\
&= \int p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{z}, \mathbf{e}, \mathbf{d}) p(\mathbf{e}, \mathbf{d}) d\mathbf{e} d\mathbf{d} && \text{By definition of the data generating process}
\end{aligned}
$$

$$= \int p(\boldsymbol{\epsilon} \mid \mathbf{z}, \mathbf{e}) p(\boldsymbol{\delta}\boldsymbol{\epsilon} \mid \mathbf{z}, \mathbf{d}) p(\mathbf{e}, \mathbf{d}) d\mathbf{e} d\mathbf{d} \qquad \text{By eq. (16)}$$

$$= \int p(\boldsymbol{\epsilon} \mid \mathbf{e}) p(\boldsymbol{\delta} \mid \mathbf{d}) p(\mathbf{e}, \mathbf{d}) d\mathbf{e} d\mathbf{d} \qquad \text{By eqs. (14) and (15)}$$

$$\begin{aligned}
p(\boldsymbol{\epsilon}, \boldsymbol{\delta}) &= \int p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{z}, \mathbf{e}, \mathbf{d}) p(\mathbf{z}, \mathbf{e}, \mathbf{d}) d\mathbf{z} d\mathbf{e} d\mathbf{d} && \text{By marginalization} \\
&= \int p(\boldsymbol{\epsilon} \mid \mathbf{z}, \mathbf{e}) p(\boldsymbol{\delta} \mid \mathbf{z}, \mathbf{d}) p(\mathbf{z}, \mathbf{e}, \mathbf{d}) d\mathbf{z} d\mathbf{e} d\mathbf{d} && \text{By eq. (16)} \\
&= \int p(\boldsymbol{\epsilon} \mid \mathbf{e}) p(\boldsymbol{\delta} \mid \mathbf{d}) p(\mathbf{z}, \mathbf{e}, \mathbf{d}) d\mathbf{z} d\mathbf{e} d\mathbf{d} && \text{By eqs. (14) and (15)} \\
&= \int p(\boldsymbol{\epsilon} \mid \mathbf{e}) p(\boldsymbol{\delta} \mid \mathbf{d}) p(\mathbf{e}, \mathbf{d} \mid \mathbf{z}) p(\mathbf{z}) d\mathbf{z} d\mathbf{e} d\mathbf{d} \\
&= \int p(\boldsymbol{\epsilon} \mid \mathbf{e}) p(\boldsymbol{\delta} \mid \mathbf{d}) p(\mathbf{e}, \mathbf{d}) p(\mathbf{z}) d\mathbf{z} d\mathbf{e} d\mathbf{d} && \text{By definition of the data generating process} \\
&= \int p(\boldsymbol{\epsilon} \mid \mathbf{e}) p(\boldsymbol{\delta} \mid \mathbf{d}) p(\mathbf{e}, \mathbf{d}) \left(\int p(\mathbf{z}) d\mathbf{z} \right) d\mathbf{e} d\mathbf{d} \\
&= \int p(\boldsymbol{\epsilon} \mid \mathbf{e}) p(\boldsymbol{\delta} \mid \mathbf{d}) p(\mathbf{e}, \mathbf{d}) d\mathbf{e} d\mathbf{d}
\end{aligned}$$

Using all of the above facts, it follows that $\boldsymbol{\delta} \perp\!\!\!\perp \boldsymbol{\epsilon} \Rightarrow \boldsymbol{\delta} \perp\!\!\!\perp \boldsymbol{\epsilon} \mid \mathbf{z}$:

$$\begin{aligned}
p(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{z}) &= p(\boldsymbol{\epsilon}, \boldsymbol{\delta}) \\
&= p(\boldsymbol{\epsilon}) p(\boldsymbol{\delta}) && \text{Since } \boldsymbol{\delta} \perp\!\!\!\perp \boldsymbol{\epsilon} \\
&= p(\boldsymbol{\epsilon} \mid \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{z}) && \text{By eqs. (12) and (13),}
\end{aligned}$$

thus satisfying the sequence of implications in eq. (11).

We have thus far established that under $\mathcal{H}_1$, $\boldsymbol{\delta} \not\perp\!\!\!\perp \boldsymbol{\epsilon}$, but under $\mathcal{H}_0$, $\boldsymbol{\delta} \perp\!\!\!\perp \boldsymbol{\epsilon}$. Therefore, given a marginal independence test $\psi(\mathcal{D}_{\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\delta}}}, \alpha) : (\mathbb{R} \times \mathbb{R})^N \times [0, 1] \to \{0, 1\}$ that is known to achieve power greater than level α, using ψ with a dataset of samples from $\hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$ will result in a conditional test with power greater than α.

□

A.5 Proof of prop. 2

Proposition 2. *Let $\boldsymbol{\epsilon} = F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z})$ and $\boldsymbol{\delta} = F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z})$. For data generating processes where both $F_{\cdot|\mathbf{z}}(\cdot \mid \mathbf{z})$ functions are invertible in the first argument and $(\boldsymbol{\epsilon}, \boldsymbol{\delta}) \perp\!\!\!\perp \mathbf{z}$,* DIET *with the following mutual information-based marginal dependence measure ρ is the most powerful conditionally valid test:*

$$\rho(\mathcal{D}_{\boldsymbol{\delta}, \boldsymbol{\epsilon}}) = \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(\boldsymbol{\delta}_i, \boldsymbol{\epsilon}_i)}{p(\boldsymbol{\delta}_i) p(\boldsymbol{\epsilon}_i)}.$$

Proof. Using the conditional CDFs $F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z})$ and $F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z})$, define the following terms for convenience:

$$\begin{aligned}
\bar{f}_{\mathbf{z}}(\mathbf{x}) &:= F_{\mathbf{x}|\mathbf{z}}(\mathbf{x} \mid \mathbf{z}) \\
\bar{g}_{\mathbf{z}}(\mathbf{y}) &:= F_{\mathbf{y}|\mathbf{z}}(\mathbf{y} \mid \mathbf{z}) \\
J &= \begin{bmatrix} \frac{d}{d\mathbf{x}} \bar{f}_{\mathbf{z}}(\mathbf{x}) & \frac{d}{d\mathbf{y}} \bar{f}_{\mathbf{z}}(\mathbf{x}) \\ \frac{d}{d\mathbf{x}} \bar{g}_{\mathbf{z}}(\mathbf{y}) & \frac{d}{d\mathbf{y}} \bar{g}_{\mathbf{z}}(\mathbf{y}) \end{bmatrix} \\
&= \begin{bmatrix} \bar{f}'_{\mathbf{z}}(\mathbf{x}) & 0 \\ 0 & \bar{g}'_{\mathbf{z}}(\mathbf{y}) \end{bmatrix}.
\end{aligned}$$

The off-diagonals of J are 0 because $\bar{f}_{\mathbf{z}}(\mathbf{x})$ is not a function of $\mathbf{y}$ and $\bar{g}_{\mathbf{z}}(\mathbf{y})$ is not a function of $\mathbf{x}$. Then using change of variables, we can write:

$$p(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = p(\boldsymbol{\epsilon} = \bar{f}_{\mathbf{z}}(\mathbf{x}), \boldsymbol{\delta} = \bar{g}_{\mathbf{z}}(\mathbf{y}) \mid \mathbf{z}) \cdot |\det(J)|$$

$$
\begin{aligned}
&= p(\boldsymbol{\epsilon} = \bar{f}_{\mathbf{z}}(\mathbf{x}), \boldsymbol{\delta} = \bar{g}_{\mathbf{z}}(\mathbf{y}) \mid \mathbf{z}) \cdot |\bar{f}'_{\mathbf{z}}(\mathbf{x}) \cdot \bar{g}'_{\mathbf{z}}(\mathbf{y})| \\
&= p(\boldsymbol{\epsilon} = \bar{f}_{\mathbf{z}}(\mathbf{x}), \boldsymbol{\delta} = \bar{g}_{\mathbf{z}}(\mathbf{y}) \mid \mathbf{z}) \cdot \bar{f}'_{\mathbf{z}}(\mathbf{x}) \cdot \bar{g}'_{\mathbf{z}}(\mathbf{y}) \\
&= p(\boldsymbol{\epsilon} = \bar{f}_{\mathbf{z}}(\mathbf{x}), \boldsymbol{\delta} = \bar{g}_{\mathbf{z}}(\mathbf{y})) \cdot \bar{f}'_{\mathbf{z}}(\mathbf{x}) \cdot \bar{g}'_{\mathbf{z}}(\mathbf{y}).
\end{aligned}
$$

The second last step follows because $\bar{f}_{\mathbf{z}}$ and $\bar{g}_{\mathbf{z}}$ are monotonically non-decreasing, meaning their derivatives with respect to $\mathbf{x}$ or $\mathbf{y}$ for a fixed $\mathbf{z}$ are non-negative. The absolute value of the product of two non-negative quantities is just the product of the two quantities. The last step uses the assumption that $(\boldsymbol{\epsilon}, \boldsymbol{\delta}) \perp\!\!\!\perp \mathbf{z}$. Using similar reasoning,

$$
\begin{aligned}
p(\mathbf{x} \mid \mathbf{z}) &= p(\boldsymbol{\epsilon} = \bar{f}_{\mathbf{z}}(\mathbf{x}) \mid \mathbf{z}) \cdot \bar{f}'_{\mathbf{z}}(\mathbf{x}) = p(\boldsymbol{\epsilon} = \bar{f}_{\mathbf{z}}(\mathbf{x})) \cdot \bar{f}'_{\mathbf{z}}(\mathbf{x}) \\
p(\mathbf{y} \mid \mathbf{z}) &= p(\boldsymbol{\delta} = \bar{g}_{\mathbf{z}}(\mathbf{y}) \mid \mathbf{z}) \cdot \bar{g}'_{\mathbf{z}}(\mathbf{y}) = p(\boldsymbol{\delta} = \bar{g}_{\mathbf{z}}(\mathbf{y})) \cdot \bar{g}'_{\mathbf{z}}(\mathbf{y}).
\end{aligned}
$$

Using the above change of variable results, we can manipulate the likelihood ratio statistic that Katsevich and Ramdas (2020) prove is the conditionally most powerful against point alternatives.

$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \log \frac{p(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)}, \mathbf{z}^{(i)})}{p(\mathbf{y}^{(i)} \mid \mathbf{z}^{(i)})} &= \frac{1}{N}\sum_{i=1}^{N} \log \frac{p(\mathbf{x}^{(i)}, \mathbf{y}^{(i)} \mid \mathbf{z}^{(i)})}{p(\mathbf{x}^{(i)} \mid \mathbf{z}^{(i)}) p(\mathbf{y}^{(i)} \mid \mathbf{z}^{(i)})} \\
&= \frac{1}{N}\sum_{i=1}^{N} \log \frac{p(\boldsymbol{\epsilon} = \bar{f}_{\mathbf{z}^{(i)}}(\mathbf{x}^{(i)}), \boldsymbol{\delta} = \bar{g}_{\mathbf{z}^{(i)}}(\mathbf{y}^{(i)})) \cdot \bar{f}'_{\mathbf{z}^{(i)}}(\mathbf{x}^{(i)}) \cdot \bar{g}'_{\mathbf{z}^{(i)}}(\mathbf{y}^{(i)})}{p(\boldsymbol{\epsilon} = \bar{f}_{\mathbf{z}}(\mathbf{x})) \cdot \bar{f}'_{\mathbf{z}}(\mathbf{x}) \cdot p(\boldsymbol{\delta} = \bar{g}_{\mathbf{z}}(\mathbf{y})) \cdot \bar{g}'_{\mathbf{z}}(\mathbf{y})} \\
&= \frac{1}{N}\sum_{i=1}^{N} \log \frac{p(\boldsymbol{\epsilon} = \bar{f}_{\mathbf{z}^{(i)}}(\mathbf{x}^{(i)}), \boldsymbol{\delta} = \bar{g}_{\mathbf{z}^{(i)}}(\mathbf{y}^{(i)}))}{p(\boldsymbol{\epsilon} = \bar{f}_{\mathbf{z}}(\mathbf{x})) \cdot p(\boldsymbol{\delta} = \bar{g}_{\mathbf{z}}(\mathbf{y}))}.
\end{aligned}
$$

Note that this final term on is exactly the mutual-information based marginal dependence measure in the statement of prop. 2. Therefore, the optimal DIET solution is the most powerful conditionally valid test against point alternatives.

□

A.6 DCRTs satisfies conditions in theorem 2 for additive data generating processes

Recall the conditions on $\hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$ in theorem 2 required for a conditional independence test to have power greater than its size:

$$
\begin{aligned}
\hat{q}(\boldsymbol{\epsilon}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) &= p(\boldsymbol{\epsilon} \mid \mathbf{x}, \mathbf{z}) p(\boldsymbol{\delta} \mid \mathbf{y}, \mathbf{z}), && \text{(factorization)} \\
\exists \tilde{f}, \tilde{g} \quad \text{s.t.} \quad \mathbf{x} \overset{\text{a.s.}}{=} \tilde{f}(\boldsymbol{\epsilon}, \mathbf{z}), \quad \text{and} \quad \mathbf{y} \overset{\text{a.s.}}{=} \tilde{g}(\boldsymbol{\delta}, \mathbf{z}), && \text{(reconstruction)} \\
(\mathbf{d}, \boldsymbol{\delta}) \perp\!\!\!\perp \mathbf{z} \qquad (\mathbf{e}, \boldsymbol{\epsilon}) \perp\!\!\!\perp \mathbf{z}. && \text{(joint independence)}
\end{aligned}
$$

We will show that the d_0-CRT, which reduces a conditional test of independence to a marginal one, satisfies the conditions in theorem 2 under additive noise assumptions on the data generating process. The d_I-CRT is not discussed here, as it does not reduce $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$ to marginal test of independence between two univariate quantities $\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{\delta}$.

Recall that the d_0-CRT computes the marginal independence between:

$$
\boldsymbol{\epsilon} = \mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{z}] \text{ and } \boldsymbol{\delta} = \mathbf{y} - \mathbb{E}[\mathbf{y} \mid \mathbf{z}].
$$

Since $\boldsymbol{\epsilon}$ only depends on $(\mathbf{x}, \mathbf{z})$ and $\boldsymbol{\delta}$ only depends on $(\mathbf{y}, \mathbf{z})$, factorization is satisfied. Next, consider an additive generating process

$$
\begin{aligned}
\mathbf{x} &= \bar{f}(\mathbf{z}) + \mathbf{e} \\
\mathbf{y} &= \bar{g}(\mathbf{z}) + \mathbf{d},
\end{aligned}
$$

where $\mathbf{e}, \mathbf{d} \perp\!\!\!\perp \mathbf{z}$ Without loss of generality, $\mathbf{e}$ and $\mathbf{d}$ can have zero expectation. Under the assumption of additivity, $\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{z}] = \mathbf{x} - \bar{f}(\mathbf{z}) = \mathbf{e}$ and $\mathbf{y} - \mathbb{E}[\mathbf{y} \mid \mathbf{z}] = \mathbf{y} - \bar{g}(\mathbf{z}) = \mathbf{d}$. Therefore, computing $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ recovers $\mathbf{e}$ and $\mathbf{d}$ exactly.

Recovering $\mathbf{e}$ and $\mathbf{d}$ exactly also implies that reconstruction and joint independence are satisfied. This is because with knowledge of $\mathbf{z}$ and $\mathbf{e}$, $\mathbf{x}$ can be reconstructed exactly. The same holds for $\mathbf{y}$ with $\mathbf{z}$ and $\mathbf{d}$. The joint independence property holds trivially because $\boldsymbol{\epsilon} = \mathbf{e}$ and $\boldsymbol{\delta} = \mathbf{d}$, and $\mathbf{e}$ and $\mathbf{d}$ are independent of $\mathbf{z}$ by definition.

Thus, all three conditions in theorem 2 are satisfied by the d_0-CRT under additivity.

B EXPERIMENTAL DETAILS

All experiments are run using a single Intel Xeon Platinum 8268 2.9GHz CPU and an NVIDIA RTX8000 GPU.

B.1 Training and hyperparameter details for DIET with MDNs

To fit the MDNs, we use the following network architecture to model each of $\hat{Q}_{\text{CDF}(\mathbf{y}|\mathbf{z})}(\mathbf{y} \mid \mathbf{z}; \theta)$ and $\hat{Q}_{\text{CDF}(\mathbf{x}|\mathbf{z})}(\mathbf{x} \mid \mathbf{z}; \eta)$. We give details about modelling $\hat{Q}_{\text{CDF}(\mathbf{y}|\mathbf{z})}(\mathbf{y} \mid \mathbf{z}; \theta)$, but the model for $\hat{Q}_{\text{CDF}(\mathbf{x}|\mathbf{z})}(\mathbf{x} \mid \mathbf{z}; \eta)$ is identical. The network consists of six consecutive fully-connected layers each followed by batch normalization and ReLU activation. For each input $\mathbf{z}^{(i)}$, the neural network outputs mixture parameters π_θ, mean parameters μ_θ, and variance parameters σ_θ, each consisting of K dimensions. Then, the log-likelihood of $\mathbf{y}^{(i)} \mid \mathbf{z}^{(i)}$ is computed as:

$$\log \sum_{k=1}^{K} \pi_\theta^{(k)} \mathcal{N}(\mathbf{y}^{(i)}; \mu_\theta^{(k)}, \sigma_\theta^{(k)}).$$

The training objective involves maximizing average of this quantity over all samples $(\mathbf{z}^{(i)}, \mathbf{y}^{(i)})$ in the dataset with respect to the parameters $\theta := \{\pi_\theta, \mu_\theta, \sigma_\theta\}$. This is shown in eq. (3). Letting Φ be the CDF of a standard normal random variable, the empirical CDF implied by parameters θ evaluated at a point $(\mathbf{z}^{(i)}, \mathbf{y}^{(i)})$ is:

$$\hat{Q}_{\text{CDF}(\mathbf{y}|\mathbf{z})}(\mathbf{y} = \mathbf{y}^{(i)} \mid \mathbf{z} = \mathbf{z}^{(i)}; \theta) = \sum_{k=1}^{K} \pi_\theta^{(k)} \Phi\left(\frac{\mathbf{y}^{(i)} - \mu_\theta^{(k)}}{\sigma_\theta^{(k)}}\right).$$

We employ the Adam (Kingma and Ba, 2014) optimizer with an initial learning rate of 1×10^{-3}. In our experiments, we fix $K = 10$. Our choice of marginal dependence statistic ρ discretizes $\hat{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{\delta}}$, then applies the adjusted mutual information estimator from Vinh et al. (2009).

B.2 Training and hyperparameter details for baseline CRTs

Test statistic for d_0-CRT. In this section, we review the full p-value computation for d_0-CRTs. We implement the Lasso-based models prescribed by Liu et al. (2020). This involves first fitting two regressions with ℓ_1 regularization:

$$\arg\min_\theta \sum_{i=1}^{N} (\mathbf{y}^{(i)} - \mathbf{z}^{(i)}\theta)^2 + \lambda_\theta ||\theta||_1, \quad \arg\min_\eta \sum_{i=1}^{N} (\widetilde{\mathbf{x}}^{(i)} - \mathbf{z}^{(i)}\eta)^2 + \lambda_\eta ||\eta||_1.$$

The regularization coefficients λ_θ and λ_η are found using 5-fold cross-validation. The test statistics $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}})$ and $T(\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)})$ are computed as follows:

$$T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}) = \left(\frac{\sum_{i=1}^{N} (\mathbf{y}^{(i)} - \mathbf{z}^{(i)}\theta)(\mathbf{x}^{(i)} - \mathbf{z}^{(i)}\eta)}{\sum_{i=1}^{N} (\mathbf{x}^{(i)} - \mathbf{z}^{(i)}\eta)^2}\right)^2$$
$$T(\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)}) = \left(\frac{\sum_{i=1}^{N} (\mathbf{y}^{(i)} - \mathbf{z}^{(i)}\theta)(\widetilde{\mathbf{x}}^{(i,m)} - \mathbf{z}^{(i)}\eta)}{\sum_{i=1}^{N} (\widetilde{\mathbf{x}}^{(i,m)} - \mathbf{z}^{(i)}\eta)^2}\right)^2,$$

where $\widetilde{\mathbf{x}}^{(i,m)}$ is the ith sample of $\widetilde{\mathbf{x}}$ in $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$. Finally, the p-value for the d_0-CRT is computed as:

$$\frac{1}{M+1}\left(1 + \sum_{m=1}^{M} \mathbb{1}\left(T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}) \leq T(\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)})\right)\right).$$

Test statistic for d_I-CRT. In this section, we review the full p-value computation for d_I-CRTs. We implement the method used in Liu et al. (2020). First, the following regressions are fit:

$$\arg\min_\theta \sum_{i=1}^{N} (\mathbf{y}^{(i)} - \mathbf{z}^{(i)}\theta)^2 + \lambda_\theta ||\theta||_1, \quad \arg\min_\eta \sum_{i=1}^{N} (\widetilde{\mathbf{x}}^{(i)} - \mathbf{z}^{(i)}\eta)^2 + \lambda_\eta ||\eta||_1.$$

The regularization coefficients λ_θ and λ_η are found using 5-fold cross-validation.

The test statistic $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}})$ is computed in the following manner. First, the "top k" dimensions in $\mathbf{z}$ are selected using a Lasso heuristic. Let the set of the top k dimensions be called S_k. The dimensions of $\mathbf{z}$ in S_k are those with the highest corresponding $|\theta_j|$, where θ_j is the jth coordinate of θ. The d_I-CRT then fits a model from $(\mathbf{x} - d_\mathbf{x}, d_\mathbf{y}, \mathbf{z}_{\text{top}(k)})$ to $\mathbf{y}$. To explicitly involve first-order interactions, the d_I-CRT we implement includes interaction terms between $(\mathbf{x} - d_\mathbf{x})$ and each $\mathbf{z}_j \in \mathbf{z}_{\text{top}(k)}$. Using these interaction terms, the following regression is fit:

$$\underset{\beta,\{\beta_j\}_{j\in S_k}}{\arg\min} \sum_{i=1}^{N} \left((\mathbf{y}^{(i)} - \mathbf{z}^{(i)}\theta) - \beta(\mathbf{x}^{(i)} - \mathbf{z}^{(i)}\eta) - \sum_{j\in S_k} \beta_j \mathbf{z}_j^{(i)} (\mathbf{x}^{(i)} - \mathbf{z}^{(i)}\eta) \right)^2 + \lambda(|\beta| + \sum_j |\beta_j|).$$

The ℓ_1 penalty coefficient λ is chosen through cross validation from amongst $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^{-3}, 10^2, 10^3\}$. Finally, $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}) := \beta^2 + \frac{1}{k}\sum_{j\in S_k} \beta_j^2$. This second regression is fit during each evaluation of the test statistic on dataset $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$.

The test statistic $T(\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)})$ is computed identically, but with samples from $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)}$ instead. The p-value is computed in the same way as the d_0-CRT. Since the Lasso heuristic requires a choice of hyperparameter k, we use $k = 2\log d_\mathbf{z}$, where $d_\mathbf{z}$ is the number of coordinates in $\mathbf{z}$, as recommended by Liu et al. (2020).

Test statistic for HRT. In this section, we review the full p-value computation for the HRTs used in our experiments. We use the cross-validated HRT from Tansey et al. (2022), who show it achieves higher power than the standard HRT. First, the dataset $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ is split in half into a train and test set: $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}{}^{(\text{train})}$ and $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}{}^{(\text{test})}$. The null datasets $\{\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m,\text{train})}\}_{m=1}^M$ are correspondingly split into sets $\{\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m,\text{train})}\}_{m=1}^M$ and $\{\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m,\text{test})}\}_{m=1}^M$. Then, the model $\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$, a neural network in this case, is fit using $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}{}^{(\text{train})}$. We use the same training setup as with the MDNs in DIET. P-values are then computed using only the test sets.

To compute $T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}{}^{(\text{test})})$, we let:

$$T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}{}^{(\text{test})}) = \frac{1}{N/2}\sum_{i=1}^{N} \mathcal{L}(\hat{q}_{\text{model}}, \mathbf{y}_{\text{test}}^{(i)}, \mathbf{x}_{\text{test}}^{(i)}, \mathbf{z}_{\text{test}}^{(i)}),$$

where $\mathcal{L}$ is a loss function evaluated using $\hat{q}_{\text{model}}$ and a sample from $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}{}^{(\text{test})}$. When response $\mathbf{y}$ is a continuous random variable:

$$\mathcal{L}(\hat{q}_{\text{model}}, \mathbf{y}_{\text{test}}^{(i)}, \mathbf{x}_{\text{test}}^{(i)}, \mathbf{z}_{\text{test}}^{(i)}) = (\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)})^2,$$

where $\hat{\mathbf{y}}^{(i)}$ is the predicted value of $\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x} = \mathbf{x}_{\text{test}}^{(i)}, \mathbf{z} = \mathbf{z}_{\text{test}}^{(i)})$. If $\mathbf{y}$ is discrete, the loss function is the log-probability of observing $\mathbf{y}$ given $\mathbf{x}$ and $\mathbf{z}$:

$$\mathcal{L}(\hat{q}_{\text{model}}, \mathbf{y}_{\text{test}}^{(i)}, \mathbf{x}_{\text{test}}^{(i)}, \mathbf{z}_{\text{test}}^{(i)}) = \log \hat{q}_{\text{model}}(\mathbf{y} = \mathbf{y}_{\text{test}}^{(i)} \mid \mathbf{x} = \mathbf{x}_{\text{test}}^{(i)}, \mathbf{z} = \mathbf{z}_{\text{test}}^{(i)}).$$

The null statistic $T(\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(\text{test})})$ is computed in a similar way with the same $\hat{q}_{\text{model}}$.

Next, a p-value, $\hat{\mathbf{p}}_1$, of the HRT is computed by

$$\frac{1}{M+1}\left(1 + \sum_{m=1}^{M} \mathbb{1}\left(T(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}{}^{(\text{test})}) \geq T(\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m,\text{test})})\right)\right).$$

Finally, to compute a cross-validated p-value using the HRT, we repeat all the steps above to obtain another p-value $\hat{\mathbf{p}}_2$, but exchanging the roles of the train and test sets. These two p-values $\hat{\mathbf{p}}_1$ and $\hat{\mathbf{p}}_2$ are combined by taking $\min(1, 2 \cdot \min(\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2))$.

B.3 Variable selection experimental details

In this section, we provide specific implementation details for our variable selection experiments. First recall the setup for these experiments. Given a set of d covariates $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_d\}$ and a response $\mathbf{y}$, we test the conditional independence of each coordinate $\mathbf{x}_j$ with $\mathbf{y}$ having observed all other coordinates of $\mathbf{x}_{-j}$. For simplicity, we focus on the CI test for only

a single coordinate $\mathbf{x}_j$ in this section. The procedure for the other coordinates is identical. We refer to $\mathbf{x}_j$ as $\mathbf{x}$ and $\mathbf{x}_{-j}$ as $\mathbf{z}$.

Every CRT method assumes the ability to sample from $p(\mathbf{x} \mid \mathbf{z})$ but in some of our experiments we do not allow access to this distribution. DIET with MDNs can directly model $p(\mathbf{x} \mid \mathbf{z})$, so its approximation can be used to sample null datasets $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$. However, neither the DCRTs nor the HRT have this facility. For these models, we use deep generative models to sample from $p(\mathbf{x} \mid \mathbf{z})$ (Romano et al., 2020; Sudarshan et al., 2020; Jordon et al., 2018).

Romano et al. (2020) train a generative model $\hat{q}_{\text{knockoff}}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{z}} \mid \mathbf{x}, \mathbf{z})$ from samples of $(\mathbf{x}, \mathbf{z})$, which models $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{z}}) \mid (\mathbf{x}, \mathbf{z})$, where $\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{z}}$ are random variables that satisfy the following property:

$$[\widetilde{\mathbf{x}}, \widetilde{\mathbf{z}}, \mathbf{x}, \mathbf{z}] \overset{d}{=} [\mathbf{x}, \widetilde{\mathbf{z}}, \widetilde{\mathbf{x}}, \mathbf{z}] \overset{d}{=} [\widetilde{\mathbf{x}}, \mathbf{z}, \mathbf{x}, \widetilde{\mathbf{z}}] \overset{d}{=} [\mathbf{x}, \mathbf{z}, \widetilde{\mathbf{x}}, \widetilde{\mathbf{z}}]. \qquad \text{(swap property)}$$

The model $\hat{q}_{\text{knockoff}}$ can then be used to generate a null dataset $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$. The ith sample of $\widetilde{\mathbf{x}}$ in $\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}$ is sampled by drawing $\widetilde{\mathbf{x}}^{(i)}, \widetilde{\mathbf{z}}^{(i)}$ from $\hat{q}_{\text{knockoff}}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{z}} \mid \mathbf{x} = \mathbf{x}^{(i)}, \mathbf{z} = \mathbf{z}^{(i)})$, then discarding $\widetilde{\mathbf{z}}^{(i)}$. Due to the swap property, the sample $\widetilde{\mathbf{x}}^{(i)} \mid \mathbf{z}^{(i)} \overset{d}{=} \mathbf{x}^{(i)} \mid \mathbf{z}^{(i)}$, but is conditionally independent of $\mathbf{y}^{(i)} \mid \mathbf{z}^{(i)}$. This makes $\widetilde{\mathbf{x}}^{(i)}$ drawn from $\hat{q}_{\text{knockoff}}$ a valid null sample when used in each Model-X method's p-value computation. The null datasets $\{\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(m)}\}_{m=1}^{M}$ can be drawn the same way.

It is critical to note that if type-1 error is to be controlled using the conditions laid out by prop. 1, sample splitting is required. Since the proof of prop. 1 requires that the same function W be applied to the sequence

$$W(\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}), W(\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(1)}), \ldots, W(\mathcal{D}_{\widetilde{\mathbf{x}},\mathbf{y},\mathbf{z}}{}^{(M)}),$$

any estimator for $p(\mathbf{x} \mid \mathbf{z})$ must be fit using a separate dataset. As such, we split the dataset $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}$ into a train set $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}{}^{(\text{train})}$ and a test set $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}}{}^{(\text{test})}$. We fit models for $p(\mathbf{x} \mid \mathbf{z})$ and the HRT model $\hat{q}_{\text{model}}$ using the training set, then compute p-values using the test set.

B.4 Synthetic CVS experiments setup

In this section, we provide exact simulation details for our synthetic CVS experiments.

The $\mathbf{x}$ data is sampled as follows: $\mathbf{x} \sim \sum_{k=1}^{4} \pi_k \mathcal{N}(\mu_k \cdot \mathbf{1}, \Sigma_k)$ is a mixture of autoregressive Gaussians. Each Σ_k is a 100-dimensional covariance matrix whose (i,j)th entry is $\rho_k^{|i-j|}$. We set $(\rho_1, \rho_2, \rho_3, \rho_4) = (0.7, 0.6, 0.5, 0.4)$, $(\pi_1, \pi_2, \pi_3, \pi_4) = (0.4, 0.3, 0.2, 0.1)$, and $(\mu_1, \mu_2, \mu_3, \mu_4) = (0, 20, 40, 60)$.

The response $\mathbf{y} \mid \mathbf{x} \sim \mathcal{N}(\langle \mathbf{x}, \beta \rangle, 1)$, where β is a coefficient vector. Each non-zero element of β is drawn from $3 \cdot \text{Rademacher}(0.5)$; there are 20 non-zero elements chosen randomly in each run. These non-zero elements represent the important variables each method aims to recover.

B.5 Semi-synthetic genetics experiments setup

In this section, we provide exact simulation details for our semi-synthetic genetics experiments.

To generate each dataset $\mathcal{D}_{\mathbf{x},\mathbf{y},\mathbf{z}} \in \mathbb{R}^{963 \times 100}$, we first sample a set of genes 100 $\{\mathbf{x}_j\}_{j=1}^{100}$ from a set of 20K. Let O be the running set of genes, and S be the full set of 20K genes. The first gene $\mathbf{x}_1$ is sampled uniformly from S and added to O, and removed from S. For each $j > 1$, we apply the following procedure. A gene $\mathbf{x}_k$ is drawn uniformly from O. The correlation between $\mathbf{x}_j$ and each gene in S is computed and the top 50 strongest correlated genes F are selected. The gene $\mathbf{x}_j \sim \text{Uniform}(F)$, and is added to O and removed from S. This process is repeated until S contains 100 genes.

To sample $\mathbf{y} \mid \mathbf{x}$, we apply the following procedure defined by Liang et al. (2018). The response has four main parts: two first order terms, a second order term, and a final nonlinearity term.

$$\begin{aligned} k &\in [m/4] \\ \varphi_k^{(1)}, \varphi_k^{(2)} &\sim \mathcal{N}(1,1) \\ \varphi_k^{(3)}, \varphi_k^{(4)}, \varphi_k^{(5)}, \varphi_k^{(6)} &\sim \mathcal{N}(2,1) \end{aligned}$$

$$\mathbf{y} \mid \mathbf{x} = \epsilon + \sum_{k=1}^{m/4} \varphi_k^{(1)} \mathbf{x}_{4k-3} + \varphi_k^{(3)} \mathbf{x}_{4k-2} + \varphi_k^{(4)} \mathbf{x}_{4k-3} \mathbf{x}_{4k-2} + \varphi_k^{(5)} \tanh(\varphi_k^{(2)} \mathbf{x}_{4k-1} + \varphi_k^{(6)} \mathbf{x}_{4k}).$$

Feature	DIET	HRT	d_0-CRT	d_I-CRT	Reference(s)
Age	•	•	•	•	(a, b, c, d, e, f, g, h)
Sex	•	•	•	•	
BMI	•	•			(a, b)
Race					
Weight		•		•	
Temperature					
Heart rate	•				(a)
Smoker					
Lymphocytes count					(g, h)
Lymphocytes percent					
Days since admission	•	•		•	(g)
Respiratory rate				•	(h)
Neutrophils count					(a)
Neutrophils percent					
Eosinophils count		•	•	•	(d, g, h)
Eosinophils percent	•			•	(d)
Blood urea nitrogen	•	•	•	•	(c, d, g)
Troponin					(a, c, d, g, h)
Ferritin	•			•	(b, d, g, h)
Platelet volume	•				(b, f, h)
Platelet count					(g, h)
Creatinine					(c)
Lactate dehydrogenase					(a, g, h)
D-dimer	•	•		•	(a, c, d, e, h)
C-reactive protein	•			•	(a, b, d, g)
O2 Saturation	•	•	•	•	(a, b)
O2 device			•	•	
High O2 support	•	•	•	•	(a, g)
On room air			•	•	

Table 2: DIET with MDNs selects many medically relevant variables in the health records task, while omitting variables that provide similar but redundant information. This table shows which variables each method selects. We evaluate each CRT by comparing to variables found in well-cited medical articles: (a) (Petrilli et al., 2020), (b) (Sattar et al., 2020), (c) (Mei et al., 2020), (d) (Castro et al., 2020), (e) (Zhang et al., 2020), (f) (Zhong and Peng, 2021), (g) (Ruan et al., 2020), (h) (Zhou et al., 2020).

The variable m determines the number of important features. We set m to 20 in our experiments.

B.6 Electronic health records experiment

See table 2 for a list of features.