CONTRA: Contrarian statistics for controlled variable selection

Mukund Sudarshan[†] **Aahlad Puli**[†] **Lakshmi Subramanian**[†]
Sriram Sankararaman[*] **Rajesh Ranganath**[†‡]
[†]Courant Institute, [‡]Center for Data Science
New York University
[*]Department of Computer Science
University of California, Los Angeles

Abstract

The holdout randomization test (HRT) discovers a set of covariates most predictive of a response. Given the covariate distribution, HRTs can explicitly control the false discovery rate (FDR). However, if this distribution is unknown and must be estimated from data, HRTs can inflate the FDR. To alleviate the inflation of FDR, we propose the contrarian randomization test (CONTRA), which is designed explicitly for scenarios where the covariate distribution must be estimated from data and may even be misspecified. Our key insight is to use an equal mixture of two "contrarian" probabilistic models in determining the importance of a covariate. One model is fit with the real data, while the other is fit using the same data, but with the covariate being tested replaced with samples from an estimate of the covariate distribution. CONTRA is flexible enough to achieve a power of 1 asymptotically, can reduce the FDR compared to state-of-the-art CVS methods when the covariate distribution is misspecified, and is computationally efficient in high dimensions and large sample sizes. We further demonstrate the effectiveness of CONTRA on numerous synthetic benchmarks, and highlight its capabilities on a genetic dataset.

1 INTRODUCTION

Scientific discovery often relies on identifying a subset of covariates that are most important to a response, while controlling the number of false discoveries. These selection procedures, termed controlled variable selection (CVS) (Candes et al., 2018), pose hypothesis tests for the conditional independence of each covariate $\mathbf{x}_j$ with the response $\mathbf{y}$ given the remaining covariates $\mathbf{x}_{-j}$:

$$\mathcal{H}_0 : \mathbf{x}_j \perp \mathbf{y} \mid \mathbf{x}_{-j} \text{ vs } \mathcal{H}_1 : \mathbf{x}_j \not\perp \mathbf{y} \mid \mathbf{x}_{-j} \qquad (1)$$

The advantage of performing CVS over other variable selection methods is the explicit control of the false discovery rate (FDR). Since a ground truth set of covariates is usually unknown, scientists can specify a nominal error rate τ for selecting important covariates, and can expect no more than $\tau\%$ of their discoveries to be false.

Many widely used methods to perform CVS, however, rely on strong assumptions about the population conditional distribution $q(\mathbf{y} \mid \mathbf{x})$ to provide guarantees for FDR control (Benjamini et al., 2009; Bunea et al., 2006) in finite samples. To relax these assumptions, Candes et al. (2018) introduce the conditional randomization test (CRT) framework, which facilitates FDR control assuming access to the covariate distribution. Tansey et al. (2018a) extend CRTs with holdout randomization tests (HRTs): easy to compute and powerful CVS test statistics that use black-box models $\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x})$.

Given a training set $(\mathbf{X},\mathbf{Y})$ and a test set $(\mathbf{X}',\mathbf{Y}')$, the HRT first fits model $\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x})$ using $(\mathbf{X},\mathbf{Y})$, then generates a dataset $\widetilde{\mathbf{X}}'$ of "null" variables. The j coordinate of the ith sample of $\widetilde{\mathbf{X}}'$, $\widetilde{\mathbf{x}}_j^{(i)}$, is drawn from the population conditional distribution $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$. The HRT then compares the loss $\mathcal{L}$ of $\hat{q}_{\text{model}}$ between two datasets: (a) $(\mathbf{X}',\mathbf{Y}')$, and (b) $(\mathbf{U}_j^{(m)},\mathbf{Y}')$: a set identical to $(\mathbf{X}',\mathbf{Y}')$, but with samples of the jth covariate replaced with those from $\widetilde{\mathbf{X}}'$. To assess the importance of covariate $\mathbf{x}_j$, a p-value is computed by repeating this comparison M times with a resampled $\mathbf{U}_j^{(m)}$:

$$\frac{1}{M+1}\left(1+\sum_{m=1}^{M}\mathbb{1}\left\{\mathcal{L}(\mathbf{X}',\mathbf{Y}') \geq \mathcal{L}(\mathbf{U}_j^{(m)},\mathbf{Y}')\right\}\right). \qquad (2)$$

If this p-value is below a user-specified significance threshold, the jth covariate is deemed important for

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

$\mathbf{y}$. HRTs can guarantee finite-sample control of the FDR, without assumptions on the distribution of $\mathbf{y} \mid \mathbf{x}$, and only require that the covariates and null variables satisfy the swap property (3): for any coordinate j,

$$[\mathbf{x}_j, \mathbf{x}_{-j}] \overset{d}{=} [\widetilde{\mathbf{x}}_j, \mathbf{x}_{-j}], \tag{3}$$

where $\overset{d}{=}$ indicates equality in distribution.

In practice, a few challenges exist with the HRT. First, the choice of performance metric can affect the power of the test to select important covariates. Second, the population distribution $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$ is likely unknown and must be estimated from data. If null variables are sampled from estimated distributions, they may not satisfy the swap property (3) exactly. As a result, $\mathcal{L}(\mathbf{X}', \mathbf{Y}')$ may be consistently lower than $\mathcal{L}(\mathbf{U}_j^{(m)}, \mathbf{Y}')$, especially if $\hat{q}_{\text{model}}$ exhibits spurious dependence on a null covariate: a likely occurrence as shown by Efron (2012). As a result, the HRT tends to deflate null p-values and ultimately fails to control the number of false discoveries.

In attempt to circumvent this issue, Tansey et al. (2018a) introduce calibrated HRTs, which reweight terms in the p-value computation. These weights are learned by fitting B conditional distribution estimators $\{\hat{q}_{\text{cc}}^{(b)}(\mathbf{x}_j \mid \mathbf{x}_{-j})\}_{b=1}^{B}$, each using a different bootstrap of the data. Using these estimators, the authors weight the mth term in the p-value computation by

$$w^{(m)} = \begin{cases} \frac{\hat{q}_{\text{cc}}^{(l)}(\mathbf{x}_j \mid \mathbf{x}_{-j})}{\hat{q}_{\text{cc}}^{(1)}(\mathbf{x}_j \mid \mathbf{x}_{-j})} & \text{if } \mathcal{L}(\mathbf{x}, \mathbf{y}) < \mathcal{L}(\widetilde{\mathbf{x}}_j^{(m)}, \mathbf{x}_{-j}, \mathbf{y}) \\ \frac{\hat{q}_{\text{cc}}^{(u)}(\mathbf{x}_j \mid \mathbf{x}_{-j})}{\hat{q}_{\text{cc}}^{(1)}(\mathbf{x}_j \mid \mathbf{x}_{-j})} & \text{otherwise} \end{cases}$$

where $\hat{q}_{\text{cc}}^{(l)}$ and $\hat{q}_{\text{cc}}^{(u)}$ are the lower and upper quantiles of the B estimators respectively. Intuitively, this reweighting of the p-value computation aims to make null p-values larger and the non-null p-values smaller. However, the effectiveness of this method is diminished as the sample size increases. This is because the bootstrapped interval shrinks as sample size increases, and the lower and upper quantile estimators $\hat{q}_{\text{cc}}^{(l)}$ and $\hat{q}_{\text{cc}}^{(u)}$ will be closer to $\hat{q}_{\text{cc}}^{(1)}$, meaning $w^{(m)}$ is close to 1 and has no impact on eq. (2). So, if the $\hat{q}_{\text{cc}}$ models are misspecified, the ability of this technique to sufficiently calibrate p-values is reduced. Further, the number of estimators B is often large: Tansey et al. (2018b) set $B = 100$ in their experiments. This makes calibrated HRTs computationally expensive in high dimensions.

The issues discussed so far suggest a set of desiderata for any new CVS procedure. (1) It must be flexible enough to achieve a power of 1 asymptotically. (2) It must yield higher p-values than an HRT when the swap property in eq. (3) is violated. (3) It must be computationally efficient when performing CVS in high dimensions and large sample sizes.

Related Work. CVS methods have been in the literature for a while, but have traditionally made strong assumptions about the $q(\mathbf{y} \mid \mathbf{x})$ distribution to control FDR in finite samples (Benjamini et al., 2009; Bunea et al., 2006). To relax some of these assumptions, Candes et al. (2018) introduced the CRT. The CRT can control FDR in finite samples, but requires the generation of null variables that satisfy the swap property (3) exactly. Katsevich and Ramdas (2020) show that using the population distribution $q(\mathbf{y} \mid \mathbf{x})$ to evaluate $\mathcal{L}(\mathbf{X}', \mathbf{Y}')$ is the uniformly most powerful statistic for a CRT. In practice, they suggest fitting estimators $\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x})$ to compute this statistic. However, running CRTs with such statistics would be computationally infeasible as new models need to be fit to each null dataset sampled.

To address the generation of null variables, Bellot and van der Schaar (2019) demonstrate the utility of generative adversarial networks (GANs) in modeling each $\hat{q}_{\text{cc}}(\mathbf{x}_j \mid \mathbf{x}_{-j})$. Barber et al. (2020) provide a theoretical analysis of knockoff methods that use $\hat{q}_{\text{cc}}$ models and show that a sufficient condition for FDR control is if $\hat{q}_{\text{cc}}$ is ϵ-close in KL to $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$. However, the authors do not provide guidance on CRTs.

Liu and Janson (2020) introduce a faster version of the CRT: the distilled conditional randomization test (DCRT). The DCRT distills the information $\mathbf{x}_{-j}$ contains about $\mathbf{y}$ in a low dimensional representation to create model-based test statistics that depend only on this distilled information. This reduces the computational cost of fitting $\hat{q}_{\text{model}}$ on each null dataset sampled. However, the DCRT algorithm implicitly assumes that there are either no interactions between $\mathbf{x}_j$ and $\mathbf{x}_{-j}$ in the generating process for $\mathbf{y}$, or that the distillation process can itself be used as a heuristic measure of variable importance. The latter assumption may be problematic since the goal of CVS is to identify important covariates. Further, it relies on access to each $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$, which as discussed earlier can be an issue.

Tansey et al. (2018a) propose the HRT, which uses the performance of $\hat{q}_{\text{model}}$ on held-out data as a CRT test statistic. This procedure places no assumptions on the distribution of $\mathbf{y} \mid \mathbf{x}$, and is highly computationally efficient. Unfortunately, the performance of the HRT is severely impacted by the quality of the null variables. If null variables are estimated from finite data and do not satisfy the swap property eq. (3) exactly, HRTs deflate p-values and violate FDR control. To combat this issue, Tansey et al. (2018a) propose an HRT calibration procedure. However, the effectiveness of the method decreases with sample size as discussed earlier.

Our contribution. To address the challenges faced by CRTs and HRTs, we present the contrarian randomization test (CONTRA): a CVS procedure based on CRTs that uses the log probability of a mixture of two “contrarian” models as a CVS test statistic. This mixture consists of a model fit to the true data and a model fit to the same data, but with the jth covariate swapped out for null data. Our contributions can be summarized as follows. (1) We explore the theoretical properties of CONTRA and prove that it can control the FDR in finite samples, and that despite using “contrarian” models CONTRA achieves an asymptotic power of 1. (2) We discuss how CONTRA can yield higher p-values than the HRT when the swap property in eq. (3) is violated, improving FDR control. (3) We show that CONTRA is computationally efficient compared to calibrated HRTs and requires far fewer model evaluations. (4) We study CONTRA on several synthetic and real datasets. Across each study, CONTRA exhibits superior FDR control over multiple baselines even when the swap property is violated.

2 BACKGROUND

In this section we review holdout randomization tests (HRTs), then discuss their pitfalls in detail. Let $\mathbf{x} \in \mathbb{R}^d$ be a vector of covariates, $\mathbf{y} \in \mathbb{R}$ be a response, and $q(\mathbf{x},\mathbf{y})$ be the generating distribution over $\mathbf{x}$ and $\mathbf{y}$. Let $(\mathbf{X},\mathbf{Y}) := \{(\mathbf{x}^{(i)},\mathbf{y}^{(i)})\}_{i=1}^{n_{\text{train}}}$ be a training set of size n_{train}, and $(\mathbf{X}',\mathbf{Y}')$ be a test set of size n_{test}. Each sample $(\mathbf{x}^{(i)},\mathbf{y}^{(i)})$ in these datasets is drawn iid from $q(\mathbf{x},\mathbf{y})$. Briefly, the HRT tests the hypothesis in eq. (1) for each covariate: the conditional independence of each $\mathbf{x}_j$ with $\mathbf{y}$ having observed all other covariates $\mathbf{x}_{-j}$. Using a user-specified FDR threshold τ, it selects a set of covariates $\hat{S}$, where FDR is defined with respect to the set of null covariates S_{null}:

$$\text{FDR} := \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\frac{|\hat{S}\cap S_{\text{null}}|}{|\hat{S}|}\right].$$

HRT procedure. First, the HRT fits a model $\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x})$ using $(\mathbf{X},\mathbf{Y})$, and computes a statistic that uses $\hat{q}_{\text{model}}$'s empirical loss $\mathcal{L}$ on test set $(\mathbf{X}',\mathbf{Y}')$. Conditioned on the test set, the HRT samples M “null” datasets $\{\widetilde{\mathbf{X}}'^{(m)}\}_{m=1}^{M}$. Each dataset $\widetilde{\mathbf{X}}'^{(m)}$ consists of n_{test} samples where where the jth component of a sample $\widetilde{\mathbf{x}}^{(i)}$, $\widetilde{\mathbf{x}}_j^{(i)}$, is drawn from the conditional $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$. A set of M statistics $\{\mathcal{L}(\mathbf{U}_j^{(m)},\mathbf{Y}')\}_{m=1}^{M}$ is computed where $\mathbf{U}_j^{(m)}$ is a copy of $\mathbf{X}'$, but with the jth column swapped with that of $\widetilde{\mathbf{X}}'^{(m)}$. Finally, the importance of each $\mathbf{x}_j$ is assessed using the p-value computation in eq. (2).

Under the null hypothesis for $\mathbf{x}_j$, the swap property eq. (3) is satisfied by definition. As a result, the sequence:

$$\mathcal{T} := \{\mathcal{L}(\mathbf{U}_j^{(1)},\mathbf{Y}'),...,\mathcal{L}(\mathbf{U}_j^{(M)},\mathbf{Y}'),\mathcal{L}(\mathbf{X}',\mathbf{Y}')\}$$

is exchangeable, so p-values computed using eq. (2) stochastically will dominate a Uniform(0,1) distribution (Tansey et al., 2018a). Such p-values are sufficient to control the FDR at a nominal rate using standard multiple testing corrections like Benjamini and Yekutieli (2001) (these are summarized in appendix B).

Under the alternate hypothesis, if $\mathcal{L}(\mathbf{X}',\mathbf{Y}')$ is typically smaller than $\mathcal{L}(\mathbf{U}_j^{(m)},\mathbf{Y}')$, the HRT will yield low p-values. The advantage of this property is that the power (the probability of selecting non-null covariates) depends entirely on how well $\hat{q}_{\text{model}}$ models $q(\mathbf{y} \mid \mathbf{x})$. Using a flexible model class can yield HRTs that have an asymptotic power of 1, meaning the important covariates are never missed.

Issues with HRTs in practice. While HRTs are theoretically sound, they are not without flaw in practice. The distributions $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$ used to generate $\mathbf{U}_j^{(m)}$ are rarely known and must be estimated from data. If an estimated complete conditional $\hat{q}_{\text{cc}}(\mathbf{x}_j \mid \mathbf{x}_{-j})$ is not equal to the true conditional, the FDR of an HRT will likely be inflated.

This occurs for the following reason. Since $\hat{q}_{\text{model}}$ is fit using a finite training set, it may exhibit spurious dependence on a null covariate $\mathbf{x}_j$ due to dependence with a non-null covariate $\mathbf{x}_k$. When $\hat{q}_{\text{model}}$ is evaluated on an out-of-distribution set $(\mathbf{U}_j^{(m)},\mathbf{Y}')$, it will likely exhibit higher loss, regardless of whether or not the null hypothesis is true. In these situations, the HRT will artificially deflate the p-values computed using eq. (2). So, the HRT procedure will inflate FDR unless either $\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x}) = q(\mathbf{y} \mid \mathbf{x})$, or $\hat{q}_{\text{cc}}(\mathbf{x}_j \mid \mathbf{x}_{-j}) = q(\mathbf{x}_j \mid \mathbf{x}_{-j})$.

Tansey et al. (2018a) acknowledge this issue, and introduce a calibration procedure to correct for this behavior. However, as discussed earlier, this bootstrap-based calibration technique is ineffective in large sample sizes. Thus, it is still unclear how to leverage the HRT to yield an empirical procedure that retains the HRT's high power, produces higher p-values than the HRT despite poor $\hat{q}_{\text{cc}}$ models, and is computationally efficient.

3 CONTRARIAN STATISTICS

The primary goal of this section is to detail a procedure that is able to achieve a power of 1 asymptotically, while better controlling the FDR than HRTs when null variables must be estimated from data. We motivate CONTRA with the following intuition. The fundamental issue with HRTs is that null covariates drawn from estimated distributions can cause the loss $\mathcal{L}(\mathbf{X}',\mathbf{Y}')$ to be lower than $\mathcal{L}(\mathbf{U}_j^{(m)},\mathbf{Y}')$ even if the jth covariate is not important to $\mathbf{y}$. This is because $\hat{q}_{\text{model}}$ performs worse on the dataset $(\mathbf{U}_j^{(m)},\mathbf{Y}')$, which is not equal in distribution to $(\mathbf{X}',\mathbf{Y}')$. One solution to this problem is to bring the true and null losses closer together. By using a "contrarian" model $\hat{q}_{\text{mix}}$ – one that performs better than $\hat{q}_{\text{model}}$ on $(\mathbf{U}_j^{(m)},\mathbf{Y}')$ but worse on $(\mathbf{X}',\mathbf{Y}')$ – p-values computed using eq. (2) can be made higher. Multiple testing correction procedures will then select fewer covariates, thus lowering the FDR.

In the next few sections, we will introduce CONTRA, a procedure to build such contrarian models, then discuss its useful theoretical and empirical properties.

Building a contrarian test. Let $(\mathbf{X},\mathbf{Y})$ be a training set of size n_{train}, and $(\mathbf{X}',\mathbf{Y}')$ be a test set of size n_{test}. Each sample $(\mathbf{x}^{(i)},\mathbf{y}^{(i)})$ in these datasets is drawn iid from $q(\mathbf{x},\mathbf{y})$. CONTRA first fits a probabilistic model $\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x})$ to $(\mathbf{X},\mathbf{Y})$, and a set of conditional distribution estimators $\{\hat{q}_{\text{cc}}^{(j)}(\mathbf{x}_j \mid \mathbf{x}_{-j})\}_{j=1}^{d}$ using $\mathbf{X}$. Then, CONTRA generates $M+1$ null datasets. One to train models: $\widetilde{\mathbf{X}}$, and M to compute p-values: $\{\widetilde{\mathbf{X}}'^{(m)}\}_{m=1}^{M}$. The jth coordinate of each element $\widetilde{\mathbf{x}}^{(i)}$ in $\widetilde{\mathbf{X}}$ is drawn from the *estimated* $\hat{q}_{\text{cc}}^{(j)}(\mathbf{x}_j \mid \mathbf{x}_{-j} = \mathbf{x}^{(i)})$ conditional on the ith training sample in $\mathbf{X}$. Each $\widetilde{\mathbf{X}}'^{(m)}$ is generated the same way, but conditioned on the test set $\mathbf{X}'$ instead.

The next step in CONTRA is to fit a set of d probabilistic models $\{\hat{q}_{\text{null}}^{(j)}(\mathbf{y} \mid \widetilde{\mathbf{x}}_j,\mathbf{x}_{-j})\}_{j=1}^{d}$. Each model $\hat{q}_{\text{null}}^{(j)}$ is fit using the data $(\mathbf{U}_j,\mathbf{Y})$, where $\mathbf{U}_j$ is identical to $\mathbf{X}$, but with the jth column of $\mathbf{X}$ replaced with the jth column of $\widetilde{\mathbf{X}}$. These models will serve as the basis for our contrarian models $\{\hat{q}_{\text{mix}}^{(j)}\}_{j=1}^{d}$, where

$$\hat{q}_{\text{mix}}^{(j)}(\mathbf{y} \mid \mathbf{x}) := \frac{1}{2}\Big(\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x})+\hat{q}_{\text{null}}^{(j)}(\mathbf{y} \mid \mathbf{x})\Big).$$

Each $\hat{q}_{\text{mix}}^{(j)}$ is a mixture of the model fit to the true data $\hat{q}_{\text{model}}$, and the model fit to the null data $\hat{q}_{\text{null}}$ for the jth covariate.

To test the conditional independence of each covariate $\mathbf{x}_j$ with $\mathbf{y}$ conditioned on $\mathbf{x}_{-j}$, CONTRA first computes the following test statistic using the test set:

$$\ell^{(j)}(\mathbf{X}',\mathbf{Y}') = \sum_{i=1}^{n_{\text{test}}} -\log \hat{q}_{\text{mix}}^{(j)}(\mathbf{y} = \mathbf{y}^{(i)} \mid \mathbf{x} = \mathbf{x}^{(i)}).$$

Finally, a computation similar to eq. (2) is used to compute p-values for each covariate:

$$\frac{1}{M+1}\left(1+\sum_{m=1}^{M}\mathbb{1}\Big\{\ell^{(j)}(\mathbf{X}',\mathbf{Y}') \geq \ell^{(j)}(\mathbf{U}_j^{(m)},\mathbf{Y}')\Big\}\right) \tag{4}$$

where $\mathbf{U}_j^{(m)}$ is a copy of $\mathbf{X}'$, but with the jth column swapped with that of $\widetilde{\mathbf{X}}'^{(m)}$.

Intuitively, the use of $\hat{q}_{\text{mix}}^{(j)}$ over $\hat{q}_{\text{model}}$ will decrease the gap between $\ell^{(j)}(\mathbf{X}',\mathbf{Y}')$ and $\ell^{(j)}(\mathbf{U}_j^{(m)},\mathbf{Y}')$. This is because the mixture of $\hat{q}_{\text{model}}$ and $\hat{q}_{\text{null}}^{(j)}$ will perform worse on the set $(\mathbf{X}',\mathbf{Y}')$, but better on $(\mathbf{U}_j^{(m)},\mathbf{Y}')$. At first glance, this seems to mitigate the FDR control issue of HRTs but at the cost of power to select non-null covariates.

In the next few sections, we show that CONTRA retains the most important property of the HRT: finite sample FDR control when the null variables satisfy the swap property in eq. (3). *Despite using contrarian models*, CONTRA achieves power 1 asymptotically when the model distributions $\hat{q}_{\text{model}}$ and $\hat{q}_{\text{null}}^{(j)}$ converge in probability to $q(\mathbf{y} \mid \mathbf{x})$ and $q(\mathbf{y} \mid \mathbf{x}_{-j})$.

3.1 CONTRA controls FDR and achieves power 1

To prove properties about CONTRA's FDR and power, we discuss the p-values produced by CONTRA.

Finite sample FDR. Procedures that control the FDR require null p-values to exhibit stochastic dominance over a Uniform(0,1) random variable (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001).

Proposition 1. *Null p-values p_j produced by* CONTRA *on a test dataset $(\mathbf{X}',\mathbf{Y}')$ will stochastically dominate $\mathbf{u} \sim$ Uniform$(0,1)$ for any covariate $\mathbf{x}_j$ that is independent of response $\mathbf{y}$ having observed the other covariates $\mathbf{x}_{-j}$.*

The proof of prop. 1 can be seen from the fact that CONTRA is a variant of the CRT. We show this formally in appendix A.1, but outline a sketch here. We note that for each null covariate $\mathbf{x}_j$, the $(\mathbf{X}',\mathbf{Y}')$ is equal in distribution to any null dataset $(\mathbf{U}_j^{(m)},\mathbf{Y}')$. As a result, $\ell^{(j)}(\mathbf{X}',\mathbf{Y}')$ is equal in distribution to $\ell^{(j)}(\mathbf{U}_j^{(m)},\mathbf{Y}')$. Since the distribution functions of the test and null statistics are equal, they share the same cumulative distribution function (CDF). We can then show that the p-value computation in eq. (4) will yield a random variable whose CDF is always greater than or equal to the CDF of a uniform random variable: the definition of stochastic dominance.

Asymptotic power of 1. The power of a CVS procedure is the probability that an important covariate $\mathbf{x}_j$ is selected. An important covariate $\mathbf{x}_j$ is selected only when its p-value is below a certain threshold. Intuitively then, the lower the p-value, the more likely $\mathbf{x}_j$ is to be selected.

Proposition 2. *If $\hat{q}_{model}$ and $\hat{q}^{(j)}_{null}$ converge in probability to distributions $q(\mathbf{y} \mid \mathbf{x})$ and $q(\mathbf{y} \mid \mathbf{x}_{-j})$ respectively, the* CONTRA *p-value for an important covariate $\mathbf{x}_j$ will converge in probability to 0 in the limit of the sample size, thus yielding a method with power 1.*

The proof sketch is as follows. We know that $\hat{q}^{(j)}_{\text{null}}$ converges in probability to $q(\mathbf{y} \mid \mathbf{x}_{-j})$ since $\widetilde{\mathbf{x}}_j$ is generated specifically to be independent of $\mathbf{y} \mid \mathbf{x}_{-j}$. We then analyze the difference of the two inner terms of the p-value computation in eq. (4) by showing that $\ell^{(j)}(\mathbf{X}',\mathbf{Y}')-\ell^{(j)}(\mathbf{U}^{(m)}_j,\mathbf{Y}') < 0$. We show that an upper bound for this difference is the sum of two negative KL terms, which will be strictly negative when the null hypothesis is not true. The proof is shown in appendix A.2.

Prop. 2 highlights a noteworthy property of CONTRA. Despite using $\hat{q}^{(j)}_{\text{mix}}$, which is designed *intentionally* to exhibit higher loss than $\hat{q}_{\text{model}}$ on the test set, the asymptotic guarantees of CONTRA are just as strong as those of any HRT. While it is theoretically possible for HRTs to enjoy higher power in small sample sizes, we will soon show empirically that this difference in power is negligible.

3.2 CONTRA prevents FDR inflation

We have thus far seen that CONTRA preserves the useful attributes of HRTs: finite sample FDR control when $\hat{q}^{(j)}_{\text{cc}}(\mathbf{x}_j \mid \mathbf{x}_{-j}) = q(\mathbf{x}_j \mid \mathbf{x}_{-j})$ and an asymptotic power of 1. In this section, we will discuss the primary advantages of CONTRA over the HRT that make it a useful empirical procedure: (a) its null p-values are higher than those of the HRT when the swap property is violated, and (b) it is still computationally efficient with respect to HRTs.

Higher null p-values. To highlight the main pitfall of HRTs in practice, consider the following scenario. Let $\mathbf{x}_k$ and $\mathbf{x}_j$ be two covariates that have high mutual information, but only $\mathbf{x}_k$ is in the Markov blanket of the response $\mathbf{y}$. In finite samples, $\hat{q}_{\text{model}}$ can exhibit spurious dependence on $\mathbf{x}_j$ (Efron, 2012). As a result, if the estimated $\hat{q}^{(j)}_{\text{cc}} \neq q(\mathbf{x}_j \mid \mathbf{x}_{-j})$, the loss of $\hat{q}_{\text{model}}$ on $(\mathbf{X}',\mathbf{Y}')$ will typically be less than its loss on $(\mathbf{U}^{(m)}_j,\mathbf{Y}')$, *even when $\mathbf{x}_j$ is not important to* $\mathbf{y}$. This is because the performance of $\hat{q}_{\text{model}}$ will suffer when it is evaluated on a distribution other than the one being trained, as studied in the domain adaption literature (Crammer et al., 2008; Daumé III, 2009). In these situations, the resulting p-values will be deflated, leading to a violation of FDR control.

Contrarian models prevent deflated p-values. To understand how CONTRA does so, consider the loss of $\hat{q}^{(j)}_{\text{mix}}$ on each of $(\mathbf{X}',\mathbf{Y}')$ and $(\mathbf{U}^{(m)}_j,\mathbf{Y}')$. The mixture $\hat{q}^{(j)}_{\text{mix}}$ contains $\hat{q}^{(j)}_{\text{null}}$, which is explicitly fit to data containing samples from $\hat{q}^{(j)}_{\text{cc}}$, and thus performs better than $\hat{q}_{\text{model}}$ on $(\mathbf{U}^{(m)}_j,\mathbf{Y}')$. Additionally, the inclusion of $\hat{q}^{(j)}_{\text{null}}$ in $\hat{q}^{(j)}_{\text{mix}}$ will also result in worse performance than $\hat{q}_{\text{model}}$ on $(\mathbf{X}',\mathbf{Y}')$. Consequently, the indicator function in the CONTRA p-value computation (4) will be 1 with greater probability than the inner term of the HRT p-value (2) across datasets $(\mathbf{X}',\mathbf{Y}',\mathbf{U}^{(m)}_j)$.

A further advantage of using $\hat{q}^{(j)}_{\text{mix}}$ over $\hat{q}_{\text{model}}$ in practice is observed when the supports of $\hat{q}^{(j)}_{\text{cc}}$ and $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$ do not match. In such cases, the log-likelihood of $\hat{q}_{\text{model}}$ is not well-defined, while the log-likelihood of $\hat{q}^{(j)}_{\text{mix}}$ is well-defined. This relates to the theoretical analysis of Barber et al. (2020), who show that the empirical KL between $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$ and $\hat{q}^{(j)}_{\text{cc}}$ bounds the FDR in the case of knockoffs. See appendix D for a full discussion.

Computational efficiency. CONTRA requires an estimator $\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x})$ for $q(\mathbf{y} \mid \mathbf{x})$ fit using training data $(\mathbf{X},\mathbf{Y})$. For each covariate $\mathbf{x}_j$, it also requires a conditional model $\hat{q}^{(j)}_{\text{cc}}(\mathbf{x}_j \mid \mathbf{x}_{-j})$, and a single null model $\hat{q}^{(j)}_{\text{null}}(\mathbf{y} \mid \widetilde{\mathbf{x}}_j,\mathbf{x}_{-j})$ fit using $(\mathbf{U}_j,\mathbf{Y})$, where $\mathbf{U}_j$ is a copy of $\mathbf{X}$, but with the jth column replaced with samples from $\hat{q}^{(j)}_{\text{cc}}$. This means there are $2d{+}1$ models fit in total.

To compute p-values using these models and a test set $(\mathbf{X}',\mathbf{Y}')$, M null datasets $\{\widetilde{\mathbf{X}}'^{(m)}\}_{m=1}^{M}$ must be sampled. For each covariate, $\hat{q}_{\text{mix}}$ must be evaluated on the test sets to compute loss $\ell^{(j)}$. This results in a total of $2d{\cdot}M$ model evaluations, as there are M null replications for each of the d covariates, and $\hat{q}^{(j)}_{\text{mix}}$ consists of both $\hat{q}_{\text{model}}$ and $\hat{q}^{(j)}_{\text{null}}$. It is worthy to note, in addition, that the computations required for the jth covariate are independent of those required for all other covariates, making CONTRA embarrassingly parallel.

In comparison to CONTRA, HRTs still need to fit $d{+}1$ models ($\hat{q}_{\text{model}}$ and $\{\hat{q}^{(j)}_{\text{cc}}\}_{j=1}^{d}$), and also sample M null datasets. However, since the HRT loss only involves $\hat{q}_{\text{model}}$, a total of $d{\cdot}M$ model evaluations on the test sets are required.

Thus, CONTRA is able to lessen the FDR compared to HRTs when $\hat{q}^{(j)}_{\text{cc}} \neq q(\mathbf{x}_j \mid \mathbf{x}_{-j})$ at the cost of only a constant factor increase in the number of models fit and evaluated. This makes CONTRA a compelling method in practice.

4 EXPERIMENTS

We analyze the performance of CONTRA on several synthetic and real datasets and compare it to several well-studied CVS baselines.

Baselines. We compare CONTRA to popular CRT-based CVS methods. Recall that the $\hat{q}_{\text{model}}$-based CRT statistic discussed by Liu and Janson (2020) requires $\mathcal{O}(M)$ models to be fit for *every* covariate (Tansey et al., 2018a). This makes it highly impractical to use with model-based test statistics as discussed in this paper. As a result, we use CRTs with the computationally efficient marginal correlation statistic, which involves a p-value computation (2) using

$$\mathcal{L}(\mathbf{X}',\mathbf{Y}') = \sum_{i=1}^{n_{\text{test}}} (\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}_j)(\mathbf{y}^{(i)} - \bar{\mathbf{y}}),$$

where $\bar{\mathbf{x}}_j$ and $\bar{\mathbf{y}}$ are the sample averages of $\mathbf{x}_j$ and $\mathbf{y}$ respectively computed from the training set $(\mathbf{X},\mathbf{Y})$. We term this the CORR-CRT. For HRTs, we use two different model-based statistics:

$$\mathcal{L}_1(\mathbf{X}',\mathbf{Y}') = \sum_{i=1}^{n_{\text{test}}} -\log \hat{q}_{\text{model}}(\mathbf{y} = \mathbf{y}^{(i)} \mid \mathbf{x} = \mathbf{x}^{(i)})$$

$$\mathcal{L}_2(\mathbf{X}',\mathbf{Y}') = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{1}\{\mathbf{y}^{(i)} \neq \hat{\mathbf{y}}^{(i)})\}$$

$$\hat{\mathbf{y}}^{(i)} \sim \hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x} = \mathbf{x}^{(i)})$$

The statistic $\mathcal{L}_1$, termed the LL-HRT, is the negative log-likelihood of the test set using $\hat{q}_{\text{model}}$. The statistic $\mathcal{L}_2$, termed the 01-HRT, measures the misclassification rate of $\hat{q}_{\text{model}}$ on the test set when $\mathbf{y}$ is a discrete random variable. We exclude comparisons to calibrated HRTs, as they take many times as long to run. Fitting at least 100 $\hat{q}_{\text{cc}}$ models for every covariate, as suggested by code from Tansey et al. (2018b), proved to be significantly slower than other CVS methods for the synthetic experiments, and computationally infeasible for a high-dimensional genomics task.

4.1 Synthetic data experiments

Each experiment involving a synthetic dataset uses the following setup. First, we generate the training dataset $(\mathbf{X},\mathbf{Y})$ of n_{train} samples and a held-out test set $(\mathbf{X}',\mathbf{Y}')$ of n_{test} samples from data distribution $q(\mathbf{x},\mathbf{y})$. Each sample of covariates $\mathbf{x}^{(i)} \in \mathbb{R}^d$, and the responses $\mathbf{y}^{(i)} \in \{0,1\}$.

Next, we create d conditional models: one $\hat{q}_{\text{cc}}^{(j)}(\mathbf{x}_j \mid \mathbf{x}_{-j})$ for each $j \in \{1,...,d\}$. Since we need to be able to sample from each $\hat{q}_{\text{cc}}^{(j)}$, we implement neural histogram estimators (Miscouridou et al., 2018), which are flexible approximations to conditional densities. Each $\hat{q}_{\text{cc}}^{(j)}$ is a two-layer fully connected networks with 32 units in each layer, and a softmax output with K classes. To fit $\hat{q}_{\text{cc}}^{(j)}$, we first bin the jth column of $\mathbf{X}$ by value into K bins, then fit the neural network to predict the bin of $\mathbf{x}_j^{(i)}$ given $\mathbf{x}_{-j}^{(i)}$. Each neural network is trained with the cross-entropy loss using SGD. In our experiments, we use $K = 20$. 18 of the bins in $\hat{q}_{\text{cc}}^{(j)}$ are uniformly spaced between the 5th and 95th quantiles of each $\mathbf{x}_j$. The remaining two bins represent any samples below the 5th quantile, or above the 95th quantile. To generate samples from $\hat{q}_{\text{cc}}^{(j)}$, we use the median value of training samples in the bin that corresponds to the network's prediction given $\mathbf{x}_{-j}^{(i)}$. These models are used to generate $M+1$ null datasets $\widetilde{\mathbf{X}}$ and $\{\widetilde{\mathbf{X}}'^{(m)}\}_{m=1}^{M}$, where $\widetilde{\mathbf{X}}$ is generated conditional on $\mathbf{X}$, and each $\widetilde{\mathbf{X}}'^{(m)}$ is generated conditional on $\mathbf{X}'$. In each of our synthetic experiments, we set M to 100, unless otherwise specified.

For each of $\hat{q}_{\text{model}}$ and $\hat{q}_{\text{null}}$, we use random forests with 100 trees fit to the training set. In general, we suggest using the model, parametric or nonparametric, that performs best on a validation split of $(\mathbf{X},\mathbf{Y})$ for high power.

Finally, we compute p-values for each of CONTRA, CORR-CRT, LL-HRT, and 01-HRT. A p-value threshold is obtained using the Benjamini and Hochberg (1995) procedure to select important covariates at a pre-specified FDR. We run each experiment on a 16-core CPU with 64GB of memory.

Benchmark datasets. Our tests on four different synthetic datasets highlight differences between each CVS approach. Datasets in this section consists of $N = 2000$ samples, and $d = 20$ covariates, unless otherwise specified. We use 70% of the data as a training set to fit each $\hat{q}_{\text{cc}}^{(j)}$, $\hat{q}_{\text{model}}$, and $\hat{q}_{\text{null}}^{(j)}$. We use the remaining 30% to compute p-values.

[`orng`, `orng-c`]: As a first example, we test the case where $\mathbf{y}$ is a nonlinear function of $\mathbf{x}$, we use the `orng` and `orng-c` datasets (Chen et al., 2018). The data is generated in the following manner:

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

$$\mathbf{y} = \begin{cases} 1 & \text{if } \exp\left(\sum_{j=1}^{\ell} \mathbf{x}_j^2 - \ell\right) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

where Σ is the 20-dimensional identity in the case of `orng`. For `orng-c`, we set all off-diagonals to 0.2, and set diagonal values to 1. The variable ℓ controls the number of important covariates, which we set to 4 for both of these experiments.

[`xor`, `xor-c`]: The choice of test statistic can impact power when covariates on their own are not informative

Dataset	orng	orng-c	xor	xor-c
CONTRA	**0.95**	**1.00**	**0.97**	**0.95**
01-HRT	0.94	0.94	0.95	0.92
LL-HRT	**0.95**	0.95	0.95	0.93
CORR-CRT	0.22	0.35	0.45	0.38

Table 1: CONTRA achieves highest FCAUC ratios on synthetic data benchmarks. (Scores closer to 1 are better). While both CONTRA and the HRTs achieve similar power, the HRTs achieve worse FDP, yielding lower FCAUC ratios.

but together provide information. To explore this, we design the xor and xor-c datasets. For xor and xor-c, we first sample $\mathbf{x}$ in the same way as orng and orng-c respectively. An affine transformation is then applied to each sample, and $\mathbf{y}$ is generated in the following manner:

$$s_1,s_2 \sim 4\cdot\text{Rademacher}(0.5)$$
$$(\mathbf{x}_1,\mathbf{x}_2) \leftarrow (\mathbf{x}_1+s_1,\mathbf{x}_2+s_2)$$
$$\mathbf{y} = \begin{cases} 0 & \text{if } s_1 s_2 < 0 \\ 1 & \text{if } s_1 s_2 > 0 \end{cases}.$$

Only the first two covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ are in the Markov blanket of $\mathbf{y}$.

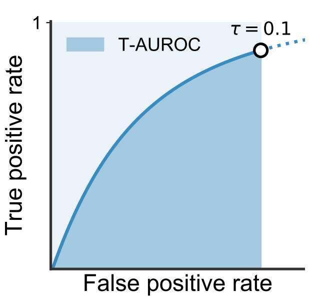

Figure 1: FDR-controlled area under the ROC curve (FCAUC) ratio: ratio of dark blue area to all blue areas.

Selection results. For each synthetic benchmark and CVS method, we run 100 experiments as described earlier to obtain p-values for each covariate. In order to concisely summarize the performance of each CVS method, we compute the FCAUC (Yu, 2012), which compute the area under a receiver operating characteristic (ROC) curve, but only up to a realistic nominal FDR. For example, practitioners are unlikely to be interested in controlling FDR at rates greater than 50%. To compute an FCAUC score, we first measure the true positive rate (TPR) (also known as power) and false positive rate (FPR) at every p-value threshold to compute a ROC curve. We then identify a nominal p-value threshold τ that corresponds to an FDR of 10% using the Benjamini and Hochberg (1995) procedure. Using the ROC curve, we compute two quantities: (A) the area under this curve from 0 to FPR(τ) (the FCAUC), and (B) the area of the rectangle defined by (0,0) and (FPR(τ),1), where FPR(τ) is the FPR corresponding to threshold τ (see fig. 1 for illustration). The score we assign to each CVS method is the ratio of (A) to (B): the FCAUC ratio. Intuitively, the closer this score is to 1, the higher the performance of a CVS method. Table 1 shows the average of this score for every CVS method and dataset across each of the 100 runs. Standard errors are omitted from table 1 as they are each fewer than four decimal places.

CONTRA achieves a higher FCAUC ratio than competing baselines. At a nominal FDR rate of 10%, the HRT methods tend to exhibit false discovery proportions (FDPs)[1] of 15% or more, while CONTRA maintains the FDP at or below 10%. It is worth noting that this difference in FDP is the main driver of CONTRA's higher performance in table 1. Both the HRT methods achieve power equal to that of CONTRA.

We further observe that the CORR-CRT performs noticeably worse than the other methods. This is likely due to its inability to model interactions between covariates when computing the test statistic, resulting in low power.

Table 1 shows promising results, as it suggests that despite using contrarian model $\hat{q}^{(j)}_{\text{mix}}$, CONTRA suffers no loss in power compared to the baselines on these four benchmarks. To understand the power lost due to contrarian models, we repeat the orng experiment at different sample sizes. These results are reported in appendix C. Having observed that the main difference between CONTRA and the baselines is primarily the control of FDR, we next explore questions that help further understand the useful FDR properties of CONTRA.

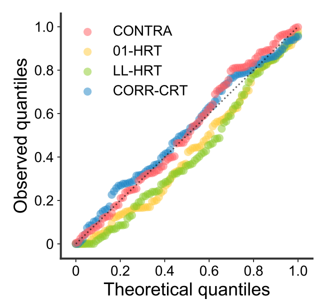

Figure 2: CONTRA exhibits null p-values that are well calibrated.

How does the choice of CVS method affect p-value calibration? The effectiveness of CVS methods to control the FDR is greatly reduced when null p-values are not super-uniform Benjamini and Hochberg (1995); Benjamini and Yekutieli (2001). For FDR to be controlled effectively, null p-values must stochastically dominate a Uniform(0,1) random variable. In this experiment, we specifically look at how well p-values produced by each CVS method satisfy this requirement for FDR control. We again use orng-c, but with one modification: we increase the number of null covariates from 16 to 100. We then perform a Kolmogorov-Smirnov hypothesis test using the set of null p-values from each CVS

[1] Another term for empirical FDR.

method. This quantifies how uniform the null p-values are.

Figure 2 shows a quantile-quantile plot of the null p-values of each method. The closer the points match the dotted black diagonal, the closer the null p-values are to Uniform(0, 1). We first notice that both CONTRA and CORR-CRT are well calibrated with Kolmogorov-Smirnov p-values of 0.183 and 0.526 respectively. However, LL-HRT and 01-HRT yield Kolmogorov-Smirnov p-values of 0.009 and 0.005 respectively. At a type-1 error threshold of 1%, both HRTs appear to yield significantly non-uniform p-values, suggesting that HRT procedures may not control the FDR well using standard multiple correction techniques. Upon closer inspection, we observe this issue as $\hat{q}_{\text{model}}$ tends to exhibit dependence on null covariates, and each $\hat{q}_{\text{cc}}^{(j)}$ is not exactly equal to the corresponding $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$. As a result, HRT test statistics tend to overestimate the importance of the null covariates, and underestimate null p-values. This is seen in fig. 2, as the observed quantiles are below the theoretical quantiles, highlighting the deflationary behavior of the null p-values. Using contrarian models protects against this behavior, as does not using a model at all in the case of CORR-CRT.

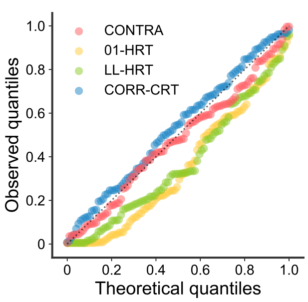

Figure 3: Despite null variable misspecification, CONTRA maintains FDR control.

What if I model $\hat{q}_{\text{cc}}$ incorrectly? In this section, we investigate the effect of modeling the null variables incorrectly on null p-values. To generate covariates, we use a mixture of autoregressive Gaussians. This provides a more challenging benchmark as each covariate is multi-modal and highly correlated with several others, encouraging $\hat{q}_{\text{model}}$ to learn spurious dependencies.

We sample $\mathbf{x} \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k \cdot \mathbf{1}, \Sigma_k)$, where each Σ_k is a 104-dimensional covariance matrix whose (i,j)th entry is $\rho_k^{|i-j|}$, and $\mathbf{1}$ is a 104-dimensional 1's vector. We set $K = 3$, and $(\rho_1,\rho_2,\rho_3) = (0.6,0.4,0.2)$. Cluster centers are set to $(\mu_1,\mu_2,\mu_3) = (0,5,10)$, and mixture proportions are set to $(\pi_1,\pi_2,\pi_3) = (0.4,0.2,0.4)$. We model all $\hat{q}_{\text{cc}}^{(j)}$ jointly with a multivariate normal (MVN) distribution. For visualization, we show two adjacent dimensions of the data and the maximum likelihood estimation (MLE) solution for the MVN in fig. 4.

We sample $\mathbf{y}$ in the same way as `orng`, using only the first four covariates as non-null. For this experiment,

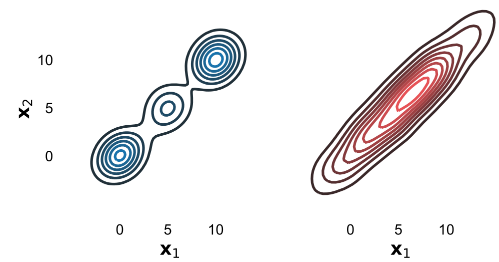

Figure 4: Data distribution: mixture of correlated Gaussians (left); Model distribution: MLE solution for multivariate Gaussian fit to data (right). Covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ are visualized.

we set the number of null resamples M to 200.

We run each CVS method on the data, and perform Kolmogorov-Smirnov hypothesis tests on the null p-values. We do not discuss the power of each method in this section, as all CVS methods other than CORR-CRT exhibit power 1 for any nominal FDR threshold above 5%. Figure 3 visualizes the null p-values for each CVS method. We observe that CONTRA and CORR-CRT both produce null p-values that appear uniform (Kolmogorov-Smirnov p-values of 0.684 and 0.399 respectively). The LL-HRT and 01-HRT produce p-values that appear to be stochastically dominated by a Uniform(0,1) random variable (Kolmogorov-Smirnov p-values of 1.059×10^{-5} and 9.342×10^{-6} respectively).

We further notice that in the range [0,0.15] on the x-axis, the HRT methods yield several p-values of 1/201, the minimum possible given our setup. Upon closer investigation, we report the following observations that explain why this p-value deflation occurs. First, $\hat{q}_{\text{model}}$ is found to exhibit spurious dependence on null covariates that correlate highly with one of $\{\mathbf{x}_i\}_{i=1}^{4}$. Second, the mixture distribution has low support on covariates in the neighborhood around (5,5), while the $\hat{q}_{\text{cc}}^{(j)}$ models place considerable mass around this point. As a result, $\hat{q}_{\text{model}}$ is evaluated on data out of its support, and consistently exhibits higher losses on the null data $(\mathbf{U}_j^{(m)},\mathbf{Y}')$ than on the test set $(\mathbf{X}',\mathbf{Y}')$, even when computing null p-values. Thus, the null p-values tend to be stochastically dominated by a Uniform(0,1) random variable and lead to the inflation of FDR.

4.2 Celiac disease experiment

Abnormalities in the genome of an individual have been found to associate with Celiac disease (Dubois et al., 2010). To understand how well CVS methods are able to replicate the results of biological studies in a purely computational procedure, we study a large genetics dataset. We apply each CVS method to a large (cases = 3.7K, controls = 8.2K) Celiac disease dataset (Dubois et al., 2010).

	# Selected	Precision	Recall	Time (s)
CONTRA	12	66.67%	20%	9207
01-HRT	15	53.33%	20%	8871
LL-HRT	14	57.14%	20%	8912
CORR-CRT	118	5.08%	15%	1512

Table 2: CONTRA achieves power on par with state-of-the-art CVS methods while achieving higher precision. Here we compare CVS methods on their ability to identify biologically relevant SNPs for Celiac disease.

In our dataset, the covariates $\mathbf{x}_j \in \{0,1,2\}$ represent single nucleotide polymorphisms (SNPs), which measure the genetic variance for each individual with respect to a reference genome. The response $\mathbf{y}$ is a binary label indicating the presence of Celiac disease. We preprocess the data as suggested by Bush and Moore (2012). First, the set of SNPs is preprocessed using linkage-disequilibrium pruning (Calus and Vandenplas, 2018), a commonly used procedure in genomics to filter out redundant SNPs using pairwise correlation. The total number of SNPs after filtering is 1759. Then, genetic principal components are added as covariates[2] to $\hat{q}_{\text{model}}$ (and $\hat{q}_{\text{null}}$) to correct for population biases (Price et al., 2006). To model $\hat{q}_{\text{cc}}$, we use the same approach as Candes et al. (2018), which uses $\hat{q}_{\text{cc}}^{(j)}$ models that condition only on a subset of SNPs in a neighborhood around $\mathbf{x}_j$, rather than all other SNPs. For exact implementation details, we refer the reader to section 7 of Candes et al. (2018). Finally, we use L_1-penalized logistic regression for $\hat{q}_{\text{model}}$ and $\hat{q}_{\text{null}}$, and set the number of null replicates M to 500.

Selection results. After running each CVS procedure on the data, we select important SNPs using a 5% FDR threshold. Using the list of SNPs returned by each method, we compare each one to the genetics literature. Specifically, we determine which SNPs have been shown to map to immunological pathways responsible for the development of Celiac disease Dubois et al. (2010); Sollid (2002); Adamovic et al. (2008); Hunt et al. (2008). If an identified SNP has been mentioned by one of these studies, we deem it important.

Table 2 shows that while CONTRA and the HRTs achieve the same recall, CONTRA achieves a higher precision (which is 1 - FDR). CORR-CRT fails to account for dependence between SNPs and tends to overestimate the variance of a single covariate, which leads to many false discoveries.

Finally, we time CONTRA and the HRTs and note that despite the high dimensionality of the problem and large M, CONTRA is only 5 minutes slower due to the fitting of $\hat{q}_{\text{null}}$ models (shown in table 2).

[2]These covariates are not tested or modeled using $\hat{q}_{\text{cc}}$.

5 DISCUSSION

CVS procedures like the HRT are popular for their ability to control the FDR. However, they can deflate p-values when the covariate distribution is unknown, thus violating FDR control. CONTRA is designed specifically for situations where the covariate distribution must be estimated from data. CONTRA is able to control FDR in finite samples, and remarkably, achieves power 1 in the limit of data despite the use of contrarian models that yield more conservative p-values than HRTs. CONTRA exhibits state-of-the-art power on several synthetic and real benchmarks, while maintaining FDR at levels closer to the nominal rate than competing baselines.

Acknowledgements

The authors would like to thank the reviewers for their thoughtful feedback. Mukund Sudarshan was partially supported by a PhRMA Foundation Predoctoral Fellowship. Mukund Sudarshan, Aahlad Puli, and Rajesh Ranganath were partly supported by NIH/NHLBI Award R01HL148248, and by NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science. Sriram Sankararaman was partially supported by NSF Award 1705121 III: Medium: Scalable Machine Learning for Genome-Wide Association Analyses.

References

Adamovic, S., Amundsen, S., Lie, B., Gudjonsdottir, A., Ascher, H., Ek, J., Van Heel, D., Nilsson, S., Sollid, L., and Naluai, Å. T. (2008). Association study of il2/il21 and fcgriia: significant association with the il2/il21 region in scandinavian coeliac disease families. *Genes and immunity*, 9(4):364.

Barber, R. F., Candes, E. J., Samworth, R. J., et al. (2020). Robust inference with knockoffs. *Annals of Statistics*, 48(3):1409–1431.

Bellot, A. and van der Schaar, M. (2019). Conditional independence testing using generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 2202–2211.

Benjamini, Y., Gavrilov, Y., et al. (2009). A simple forward selection procedure based on false discovery rate control. *The Annals of Applied Statistics*, 3(1):179–198.

Benjamini, Y. and Hochberg, Y. (1995). 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.

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

Bunea, F., Wegkamp, M. H., and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *Journal of statistical planning and inference*, 136(12):4349–4364.

Bush, W. S. and Moore, J. H. (2012). Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822.

Calus, M. P. and Vandenplas, J. (2018). Snprune: an efficient algorithm to prune large snp array and sequence datasets based on high linkage disequilibrium. *Genetics Selection Evolution*, 50(1):34.

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). 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.

Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. (2018). Learning to explain: An information-theoretic perspective on model interpretation. *arXiv preprint arXiv:1802.07814*.

Crammer, K., Kearns, M., and Wortman, J. (2008). Learning from multiple sources. *Journal of Machine Learning Research*, 9(Aug):1757–1774.

Daumé III, H. (2009). Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*.

Dubois, P. C., Trynka, G., Franke, L., Hunt, K. A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G. A., Ádány, R., Aromaa, A., et al. (2010). Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, 42(4):295.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Hunt, K. A., Zhernakova, A., Turner, G., Heap, G. A., Franke, L., Bruinenberg, M., Romanos, J., Dinesen, L. C., Ryan, A. W., Panesar, D., et al. (2008). Novel celiac disease genetic determinants related to the immune response. *Nature genetics*, 40(4):395.

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

Liu, M. and Janson, L. (2020). Fast and powerful conditional randomization testing via distillation. *arXiv preprint arXiv:2006.03980*.

Miscouridou, X., Perotte, A., Elhadad, N., and Ranganath, R. (2018). Deep survival analysis: Nonparametrics and missingness. In *Machine Learning for Healthcare Conference*, pages 244–256.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909.

Sollid, L. M. (2002). Coeliac disease: dissecting a complex inflammatory disorder. *Nature Reviews Immunology*, 2(9):647.

Tansey, W., Veitch, V., Zhang, H., Rabadan, R., and Blei, D. M. (2018a). The holdout randomization test: Principled and easy black box feature selection. *arXiv preprint arXiv:1811.00645*.

Tansey, W., Wang, Y., Blei, D. M., and Rabadan, R. (2018b). Black box fdr. *arXiv preprint arXiv:1806.03143*.

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

Yu, T. (2012). Rocs: receiver operating characteristic surface for class-skewed high-throughput data. *PloS one*, 7(7):e40598.

A PROOFS

A.1 Proof of prop. 1

We show that CONTRA preserves the finite-sample control of the FDR under the same conditions that an HRT would. In this setting, $\hat{q}_{\text{cc}}^{(j)}(\mathbf{x}_j \mid \mathbf{x}_{-j}) = q(\mathbf{x}_j \mid \mathbf{x}_{-j})$. Let $\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x})$ be fit using training dataset $(\mathbf{X},\mathbf{Y})$. Let $(\mathbf{X}',\mathbf{Y}')$ be a test set. CONTRA generates $M+1$ null datasets: $\{\widetilde{\mathbf{X}}'^{(m)}\}_{m=1}^{M}$, and $\widetilde{\mathbf{X}}$. The (i,j)th element of each $\widetilde{\mathbf{X}}'^{(m)}$ is drawn from $\hat{q}_{\text{cc}}(\widetilde{\mathbf{x}}_j \mid \mathbf{x}_j = \mathbf{x}^{(i)})$, where $\mathbf{x}^{(i)}$ is the ith sample of $\mathbf{X}'$. The dataset $\widetilde{\mathbf{X}}$ is generated the same way, but conditioning on samples from training set $\mathbf{X}$ instead. The null model for the jth covariate $\hat{q}_{\text{null}}^{(j)}(\mathbf{y} \mid \widetilde{\mathbf{x}}_j,\mathbf{x}_{-j})$ is then fit using $(\mathbf{X},\mathbf{Y})$, but with the jth column of $\mathbf{X}$ swapped out for the jth column of $\widetilde{\mathbf{X}}$. The distribution $\hat{q}_{\text{mix}}^{(j)}(\mathbf{y} \mid \mathbf{x})$, is an equal mixture of $\hat{q}_{\text{model}}$ and $\hat{q}_{\text{null}}^{(j)}$.

Recall that the jth p-value for CONTRA is

$$p_j = \frac{1}{M+1}\left(1+\sum_{m=1}^{M}\mathbb{1}\left\{\ell^{(j)}(\mathbf{X}',\mathbf{Y}') \geq \ell^{(j)}(\mathbf{U}_j^{(m)},\mathbf{Y}')\right\}\right)$$

$$\ell^{(j)}(\mathbf{X}',\mathbf{Y}') = \sum_{i=1}^{n_{\text{test}}} -\log\hat{q}_{\text{mix}}^{(j)}(\mathbf{y} = \mathbf{y}^{(i)} \mid \mathbf{x} = \mathbf{x}^{(i)}),$$

and $\mathbf{U}_j^{(m)}$ is a copy of $\mathbf{X}'$, but with the jth column swapped with that of $\widetilde{\mathbf{X}}'^{(m)}$. To show that CONTRA preserves finite-sample control of the FDR, we want to show that its null p-values are uniformly distributed.

Proof. First, we will show that the datasets $\mathbf{X}'$ and $\mathbf{U}_j^{(m)}$ are equal in distributions. Under the null, $\mathbf{x}_j$ is independent of $\mathbf{y}$ given $\mathbf{x}_{-j}$. Using (Candes et al., 2018), this means the swap property mentioned earlier in eq. (3) is satisfied:

$$[\mathbf{x}_j,\mathbf{x}_{-j}] \overset{d}{=} [\widetilde{\mathbf{x}}_j,\mathbf{x}_{-j}].$$

Since the samples in $\mathbf{X}'$ and $\mathbf{U}_j^{(m)}$ are iid, the datasets $\mathbf{X}'$ and $\mathbf{U}_j^{(m)}$ will also be equal in distribution.

Recall that the components of $\hat{q}_{\text{mix}}^{(j)}$, $\hat{q}_{\text{model}}$ and $\hat{q}_{\text{null}}^{(j)}$ are fit using the training set, and are therefore independent of $(\mathbf{X}',\mathbf{Y}')$. Therefore, the statistics $\ell^{(j)}(\mathbf{X}',\mathbf{Y}')$ and $\ell^{(j)}(\mathbf{U}_j^{(m)},\mathbf{Y}')$ are also equal in distribution.

Next, we define $F_{n_{\text{test}}}^{(j)}(t)$ to be the empirical CDF of $\ell^{(j)}(\mathbf{X}',\mathbf{Y}')$ where n_{test} is the number of samples in the test set. Note that the equality in distribution between $\mathbf{U}_j^{(m)}$ and $\mathbf{X}'$ implies that $\ell^{(j)}(\mathbf{U}_j^{(m)},\mathbf{Y}')$ has the same CDF. We also define $F_{n_{\text{test}}}^{-1}(t) = \inf\{u \in \mathbb{R} : F_{n_{\text{test}}}^{(j)}(u) \geq t\}$: the generalized inverse CDF.

Using the empirical CDF, we can represent the null p-values in a more convenient form. In the limit of M, the number of null datasets sampled, the jth CONTRA p-value is simply:

$$p_j = \mathbb{P}\left\{\ell^{(j)}(\mathbf{X}',\mathbf{Y}') \geq \ell^{(j)}(\mathbf{U}_j^{(m)},\mathbf{Y}')\right\} = F_{n_{\text{test}}}^{(j)}(\ell^{(j)}(\mathbf{U}_j^{(m)},\mathbf{Y}')).$$

With this representation of the p-value, we see that

$$\begin{aligned}\mathbb{P}\{p_j \leq \alpha\} &= \mathbb{P}\{F_{n_{\text{test}}}^{(j)}(\ell^{(j)}(\mathbf{U}_j^{(m)},\mathbf{Y}')) \leq \alpha\}\\ &= \mathbb{P}\{\ell^{(j)}(\mathbf{U}_j^{(m)},\mathbf{Y}') \leq F_{n_{\text{test}}}^{-1}(\alpha)\}\\ &= F_{n_{\text{test}}}^{(j)}(F_{n_{\text{test}}}^{-1}(\alpha)) = \alpha.\end{aligned}$$

In the first line, we use the alternative representation of the null p-value. In the second line, we apply the generalized inverse CDF to both sides. Finally, we use the definition of the CDF. Thus, the CDF of the null p-value is equal to that of a uniform random variable. In finite samples, adding 1 to the sum of M indicator functions, then dividing by $(M+1)$ ensures that p_j will stochastically dominate a Uniform(0,1) random variable since the minimum p-value is then $1/(M+1)$. □

A.2 Proof of prop. 2

For the setup of the problem, see appendix A.1. Let the model distributions $\hat{q}_{\text{model}}$ and $\hat{q}^{(j)}_{\text{null}}$ converge in probability to distributions $q(\mathbf{y} \mid \mathbf{x})$ and $q(\mathbf{y} \mid \mathbf{x}_{-j})$ respectively. We will show that the CONTRA p-values corresponding to non-null covariates will converge in probability to 0 in the limit of n_{test}, the sample size of the test set. For simplicity, we assume that all random variables are continuous, but the same reasoning holds for discrete random variables.

Proof. Let $(\mathbf{X},\mathbf{Y})$ be a training set consisting of n_{train} samples $(\mathbf{x}^{(i)},\mathbf{y}^{(i)})$. Let $\widetilde{\mathbf{X}}$ be a set of null data where the jth coordinate of its ith sample, $\widetilde{\mathbf{x}}^{(i)}_j$, is sampled from $\hat{q}^{(j)}_{\text{cc}}(\mathbf{x}_j \mid \mathbf{x}_{-j} = \mathbf{x}^{(i)}_{-j})$. Let $(\mathbf{U}_j,\mathbf{Y})$ be identical to $(\mathbf{X},\mathbf{Y})$, but with the jth coordinate swapped out for the jth coordinate of $\widetilde{\mathbf{X}}$. Now let $\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x})$ be fit using $(\mathbf{X},\mathbf{Y})$, and $\hat{q}^{(j)}_{\text{null}}$ be fit using $(\mathbf{U}_j,\mathbf{Y})$.

To formalize the assumptions in prop. 2, we use the average KL between each population distribution and its respective model distribution:

$$\lim_{n_{\text{train}}\to\infty} \mathbb{E}_{q(\mathbf{x})}\mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \log \frac{q(\mathbf{y} \mid \mathbf{x})}{\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x})} = 0$$

$$\lim_{n_{\text{train}}\to\infty} \mathbb{E}_{q(\mathbf{x}_{-j})}\mathbb{E}_{\hat{q}_{\text{cc}}(\mathbf{x}_j|\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{y}|\mathbf{x}_{-j})} \log \frac{q(\mathbf{y} \mid \mathbf{x}_{-j})}{\hat{q}^{(j)}_{\text{null}}(\mathbf{y} \mid \widetilde{\mathbf{x}}_j,\mathbf{x}_{-j})} = 0 \qquad \forall j \in \{1,\ldots,d\}.$$

Recall that $\hat{q}^{(j)}_{\text{null}}$ is fit using $(\mathbf{U}_j,\mathbf{Y})$, where the jth covariate is designed to have no dependence on $\mathbf{y}$. So $\hat{q}^{(j)}_{\text{null}}$, which has no dependence on $\mathbf{x}_j$, converges to $q(\mathbf{y} \mid \mathbf{x}_{-j})$ in KL.

Now, we use a result from Tsybakov (2008) that helps bound the squared Hellinger distance between two distributions using the KL between them:

$$\int \left(\sqrt{q(\mathbf{y} \mid \mathbf{x})} - \sqrt{\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x})}\right)^2 d\mathbf{y} \le \text{KL}(q(\mathbf{y} \mid \mathbf{x}) \parallel \hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x}))$$

$$\int \left(\sqrt{q(\mathbf{y} \mid \mathbf{x}_{-j})} - \sqrt{\hat{q}^{(j)}_{\text{null}}(\mathbf{y} \mid \widetilde{\mathbf{x}}_j,\mathbf{x}_{-j})}\right)^2 d\mathbf{y} \le \text{KL}(q(\mathbf{y} \mid \mathbf{x}_{-j}) \parallel \hat{q}^{(j)}_{\text{null}}(\mathbf{y} \mid \widetilde{\mathbf{x}}_j,\mathbf{x}_{-j}))$$

This means that the log densities of the model and population distributions must be equal almost surely when $(\mathbf{x},\mathbf{y}) \sim q(\mathbf{x})q(\mathbf{y} \mid \mathbf{x})$ and $\widetilde{\mathbf{x}}_j \sim q(\mathbf{x}_j \mid \mathbf{x}_{-j})$:

$$\lim_{n_{\text{train}}\to\infty} \log \frac{q(\mathbf{y} \mid \mathbf{x})}{\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x})} = 0$$

$$\lim_{n_{\text{train}}\to\infty} \log \frac{q(\mathbf{y} \mid \mathbf{x}_{-j})}{\hat{q}^{(j)}_{\text{null}}(\mathbf{y} \mid \widetilde{\mathbf{x}}_j,\mathbf{x}_{-j})} = 0.$$

Otherwise, the Hellinger distance will be positive, implying a positive KL between the model and population distributions.

Now, let $(\mathbf{X}',\mathbf{Y}')$ be a test set of covariates consisting of n_{test} samples. Let $\widetilde{\mathbf{X}}'^{(m)}$ be a null set sampled in the same way as $\widetilde{\mathbf{X}}$, but conditioned on samples from $\mathbf{X}'$, instead of $\mathbf{X}$. Let $\mathbf{U}^{(m)}_j$ be constructed the same way as $\mathbf{U}_j$, but using $\mathbf{X}'$ and $\widetilde{\mathbf{X}}'^{(m)}$ instead. In the limit of $n_{\text{train}},n_{\text{test}}$, we want to show that

$$\ell^{(j)}(\mathbf{X}',\mathbf{Y}') < \ell^{(j)}(\mathbf{U}^{(m)}_j,\mathbf{Y}').$$

If this inequality is true as $(n_{\text{train}},n_{\text{test}}) \to (\infty,\infty)$, then the indicator inside the p-value computation is always 0, leading to a p-value of 0.

Recall in the limit of n_{test}, $\ell^{(j)}(\mathbf{X}',\mathbf{Y}')$ is equal to $\mathbb{E}_{q(\mathbf{x},\mathbf{y})}\log\hat{q}_{\text{mix}}(\mathbf{y} \mid \mathbf{x})$. A similar limit exists for $\ell^{(j)}(\mathbf{U}^{(m)}_j,\mathbf{Y}')$.

Using this fact, we can expand $\ell^{(j)}(\mathbf{X}',\mathbf{Y}')$:

$$\begin{aligned}\lim_{(n_{\text{train}},n_{\text{test}})\to(\infty,\infty)}\ell^{(j)}(\mathbf{X}',\mathbf{Y}') &= -\mathbb{E}_{q(\mathbf{x},\mathbf{y})}\log 0.5(q(\mathbf{y}\mid\mathbf{x})+q(\mathbf{y}\mid\mathbf{x}_{-j}))\\
&= -\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{x}_j\mid\mathbf{y},\mathbf{x}_{-j})}\log 0.5(q(\mathbf{y}\mid\mathbf{x})+q(\mathbf{y}\mid\mathbf{x}_{-j}))\\
&= -\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{x}_j\mid\mathbf{y},\mathbf{x}_{-j})}\log 0.5\left(\frac{q(\mathbf{y}\mid\mathbf{x}_{-j})q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})}{q(\mathbf{x}_j\mid\mathbf{x}_{-j})}+q(\mathbf{y}\mid\mathbf{x}_{-j})\right)\\
&= -\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{x}_j\mid\mathbf{y},\mathbf{x}_{-j})}\log 0.5q(\mathbf{y}\mid\mathbf{x}_{-j})\left(\frac{q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})}{q(\mathbf{x}_j\mid\mathbf{x}_{-j})}+1\right)\\
&= -\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{x}_j\mid\mathbf{y},\mathbf{x}_{-j})}\log q(\mathbf{y}\mid\mathbf{x}_{-j})+\log 0.5\left(\frac{q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})}{q(\mathbf{x}_j\mid\mathbf{x}_{-j})}+1\right)\\
&= -\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\log q(\mathbf{y}\mid\mathbf{x}_{-j})-\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{x}_j\mid\mathbf{y},\mathbf{x}_{-j})}\log 0.5\left(\frac{q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})}{q(\mathbf{x}_j\mid\mathbf{x}_{-j})}+1\right).\end{aligned}$$

We now use the fact that the geometric mean is less than the arithmetic mean to upper bound $\ell^{(j)}(\mathbf{X}',\mathbf{Y}')$:

$$\begin{aligned}\ell^{(j)}(\mathbf{X}',\mathbf{Y}') &\leq -\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\log q(\mathbf{y}\mid\mathbf{x}_{-j})-\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{x}_j\mid\mathbf{y},\mathbf{x}_{-j})}\log\sqrt{\frac{q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})}{q(\mathbf{x}_j\mid\mathbf{x}_{-j})}}\\
&= -\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\log q(\mathbf{y}\mid\mathbf{x}_{-j})-0.5\cdot\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{x}_j\mid\mathbf{y},\mathbf{x}_{-j})}\log\frac{q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})}{q(\mathbf{x}_j\mid\mathbf{x}_{-j})}\\
&= -\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\log q(\mathbf{y}\mid\mathbf{x}_{-j})-0.5\cdot\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\mathrm{KL}(q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})\parallel q(\mathbf{x}_j\mid\mathbf{x}_{-j})).\end{aligned}$$

Using similar arithmetic, we now expand $\ell^{(j)}(\mathbf{U}_j^{(m)},\mathbf{Y}')$:

$$\begin{aligned}\lim_{(n_{\text{train}},n_{\text{test}})\to(\infty,\infty)}\ell^{(j)}(\mathbf{U}_j^{(m)},\mathbf{Y}') &= -\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{x}_j\mid\mathbf{x}_{-j})}\log 0.5(q(\mathbf{y}\mid\mathbf{x})+q(\mathbf{y}\mid\mathbf{x}_{-j}))\\
&= -\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{x}_j\mid\mathbf{x}_{-j})}\log q(\mathbf{y}\mid\mathbf{x}_{-j})+\log 0.5\left(\frac{q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})}{q(\mathbf{x}_j\mid\mathbf{x}_{-j})}+1\right)\\
&= -\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\mathbb{E}_{q(\mathbf{x}_j\mid\mathbf{x}_{-j})}\log q(\mathbf{y}\mid\mathbf{x}_{-j})+\log\left(\frac{0.5q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})+0.5q(\mathbf{x}_j\mid\mathbf{x}_{-j})}{q(\mathbf{x}_j\mid\mathbf{x}_{-j})}\right)\\
&= -\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\log q(\mathbf{y}\mid\mathbf{x}_{-j})-\mathbb{E}_{q(\mathbf{x}_j\mid\mathbf{x}_{-j})}\log\left(\frac{0.5q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})+0.5q(\mathbf{x}_j\mid\mathbf{x}_{-j})}{q(\mathbf{x}_j\mid\mathbf{x}_{-j})}\right)\\
&= -\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\log q(\mathbf{y}\mid\mathbf{x}_{-j})+\mathrm{KL}(q(\mathbf{x}_j\mid\mathbf{x}_{-j})\parallel 0.5q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})+0.5q(\mathbf{x}_j\mid\mathbf{x}_{-j})).\end{aligned}$$

Putting it all together:

$$\begin{aligned}\lim_{(n_{\text{train}},n_{\text{test}})\to(\infty,\infty)}\ell^{(j)}(\mathbf{U}_j^{(m)},\mathbf{Y}')-\ell^{(j)}(\mathbf{X}',\mathbf{Y}') &\geq \mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\log q(\mathbf{y}\mid\mathbf{x}_{-j})+0.5\cdot\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\mathrm{KL}(q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})\parallel q(\mathbf{x}_j\mid\mathbf{x}_{-j}))\\
&\quad-\left(\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\log q(\mathbf{y}\mid\mathbf{x}_{-j})-\mathrm{KL}(q(\mathbf{x}_j\mid\mathbf{x}_{-j})\parallel 0.5q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})+0.5q(\mathbf{x}_j\mid\mathbf{x}_{-j}))\right)\\
&= 0.5\cdot\mathbb{E}_{q(\mathbf{y},\mathbf{x}_{-j})}\mathrm{KL}(q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})\parallel q(\mathbf{x}_j\mid\mathbf{x}_{-j}))\\
&\quad+\mathrm{KL}(q(\mathbf{x}_j\mid\mathbf{x}_{-j})\parallel 0.5q(\mathbf{x}_j\mid\mathbf{x}_{-j},\mathbf{y})+0.5q(\mathbf{x}_j\mid\mathbf{x}_{-j}))\\
&> 0.\end{aligned}$$

The key takeaway here is that the difference between $\ell^{(j)}(\mathbf{U}_j^{(m)},\mathbf{Y}')$ and $\ell^{(j)}(\mathbf{X}',\mathbf{Y}')$ is lower bounded by the sum of two Kullback–Leibler divergence (KL) terms, both of which are strictly positive in the case of a non-null covariate. Thus, we have shown that in the limit of n_{test}, p-values produced by CONTRA will be 0 in distribution. This also implies the p-values converge to 0 in probability since 0 is a constant. □

B FDR AND BENJAMINI-HOCHBERG

In this section, we review multiple testing correction procedures, focusing specifically on the most popular one: the Benjamini-Hochberg (Benjamini and Hochberg, 1995) procedure.

Multiple testing motivation. For a single statistical hypothesis, if the p-value yielded by the test is below some confidence level α, then the null hypothesis is rejected resulting in a "discovery." The p-value being below α does not guarantee that the null hypothesis is not true. If the null is true, then with probability at most α, the null will be erroneously rejected.

If there are d such hypotheses being tested with the same data, and the null is true for each one, then we should expect to see $d{\cdot}\alpha$ false discoveries. Thus, the family-wise error rate (FWER) – the probability of making one or more false discoveries – increases with dimension. To control the FWER, one potential solution is to divide α by d, but in high dimensions, an extremely small p-value is required to make a discovery, yielding low power.

For a less conservative notion of false discovery control, Benjamini and Hochberg (1995) introduced the FDR: the ratio of false discoveries to total discoveries. Rather than seeking to bound the probability of making even a single error, Benjamini and Hochberg (1995) suggest that practitioners should aim to control the rate of false discoveries only in the set of rejected hypotheses.

Benjamini-Hochberg procedure. Given a set of d p-values $\{p_j\}_{j=1}^{d}$ and a nominal FDR rate τ, the Benjamini-Hochberg procedure seeks to identify a threshold for p-values, q, such that the proportion of null hypotheses rejected that have p-value less than q will not exceed τ. The procedure is as follows. First, assign a rank $r_j \in \{1,\ldots,d\}$ to each p-value in ascending order. Then, compute a critical value for each p_j:

$$c_j := \frac{r_j{\cdot}\tau}{d}.$$

Next, identify the largest p_j such that for all $p_k \leq p_j$, $p_k \leq c_k$ and reject their corresponding hypotheses.

This procedure controls the FDR for any set of p-values where the CDF of each p-value dominates the CDF of a Uniform(0,1) random variable. We refer the reader to Benjamini and Hochberg (1995) for all proofs.

C ADDITIONAL FIGURES

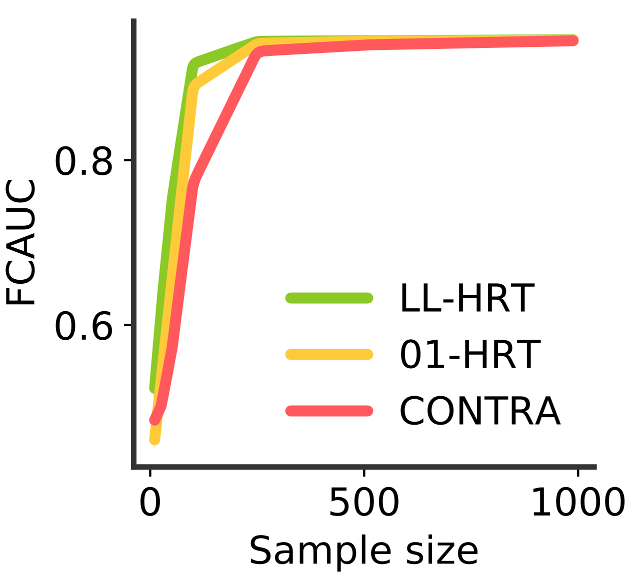

Figure 5: The loss in power due to the use of contarian models is reduced as sample size increases. Apart from very low sample sizes, CONTRA achieves power on par with both HRTs.

To understand how much power is lost due to the use of contrarian models, we repeat the `orng` experiment, but vary the sample size. We only compare CONTRA to each of the HRTs, as the power of CORR-CRT is much lower even at large sample sizes, as discussed in the experiments section. Figure 5 plots the average FCAUC for each method over all 100 replicates as a function of training sample size. The test set is equal to the training set in size. Error bars are omitted due to very small standard error.

We note there is a noticeable gap in FCAUC by CONTRA at low sample sizes (≤ 200), but minimal difference otherwise. The loss in power due to contrarian models is minimized as sample size increases. At just 500 samples, the power of CONTRA is almost equal to that of the HRTs.

D BOUNDING the FDR with KL

Background. We first provide a brief summary of knockoffs for readers unfamiliar with the concept. For a more in-depth discussion of the advantages and disadvantages of knockoffs, we refer the reader to Candes et al. (2018).

Knockoffs can be thought of as a method to compute 1-bit CRT p-values (Katsevich and Ramdas, 2020). Rather than performing a full p-value computation as shown in eq. (2), they compute a single test statistic W_j for each covariate. Let Z_j be any function of data $(\mathbf{X},\mathbf{Y})$, $\widetilde{\mathbf{X}}$ be a set of null variables sampled as in the case of a CRT, and $\mathbf{U}_j$ be identical to $\mathbf{X}$, but with the jth column swapped with that of $\widetilde{\mathbf{X}}$. The knockoff statistic W_j can be defined as:

$$W_j = Z_j(\mathbf{X},\mathbf{Y}) - Z_j(\mathbf{U}_j,\mathbf{Y}).$$

The knockoff filter selects important covariates at a nominal rate q using their respective statistics W_j by constructing an estimate of the FDR:

$$\hat{S} = \{j : W_j \geq t_0\}$$
$$t_0 = \min\left\{t > 0 : \frac{|\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}|} \leq q\right\}.$$

Requirements for FDR. The knockoff filter requires that the null variables $\widetilde{\mathbf{X}}$ are exact. That is, the jth column of the ith sample $\widetilde{\mathbf{x}}_j^{(i)}$ must be drawn from the population distribution $q(\mathbf{x}_j \mid \mathbf{x}_{-j} = \mathbf{x}_{-j}^{(i)})$, where $\mathbf{x}_{-j}^{(i)}$ is from the ith sample of $\mathbf{X}$. This is required to ensure that under the null hypothesis, $Z_j(\mathbf{X},\mathbf{Y})$ and $Z_j(\mathbf{U}_j,\mathbf{Y})$ will have the same distribution, meaning W_j is symmetric around 0. This property of W_j is termed the "flip-sign" property. Candes et al. (2018) show that when W_j satisfies the flip-sign property, the knockoff filter can control the FDR.

Issues with knockoffs in practice. Barber et al. (2020) acknowledge that practitioners seldom have access to the population distribution $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$, and must resort to estimators $\hat{q}_{\text{cc}}^{(j)}(\mathbf{x}_j \mid \mathbf{x}_{-j})$. If the population and estimated distributions are not the same, the knockoff filter may inflate the FDR beyond nominal levels. The authors show that the level to which the knockoff filter inflates the FDR is bounded by the maximum empirical KL between each $\hat{q}_{\text{cc}}^{(j)}(\mathbf{x}_j \mid \mathbf{x}_{-j})$ and $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$. They do so by relating the empirical KL to the flip-sign property. Let E_j be the empirical KL between $\hat{q}_{\text{cc}}^{(j)}(\mathbf{x}_j \mid \mathbf{x}_{-j})$ and $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$. Then for any covariate $\mathbf{x}_j$ whose null hypothesis is true,

$$\frac{\mathbb{P}(W_j > 0, E_j \leq \epsilon \mid |W_j|, W_{-j})}{\mathbb{P}(W_j < 0 \mid |W_j|, W_{-j})} \leq \exp(\epsilon) \qquad \forall \epsilon \geq 0. \tag{5}$$

We refer the reader to Barber et al. (2020) for a detailed derivation and discussion of this equation. The main takeaway is that as E_j approaches 0, the probability that W_j is positive is equal to the probability that it is negative, thus satisfying the flip-sign property required to control the FDR.

D.1 Relation of flip-sign result to CRTs

We can show how the flip-sign result of Barber et al. (2020) relates to CRTs by using the fact that W_j being symmetric around 0 implies that $Z_j(\mathbf{X},\mathbf{Y})$ and $Z_j(\mathbf{U}_j,\mathbf{Y})$ have the same distribution. A p-value using $Z_j(\mathbf{X},\mathbf{Y})$ and $Z_j(\mathbf{U}_j,\mathbf{Y})$ by performing the following computation. Letting each $\mathbf{U}_j^{(m)}$ be drawn independently the same way as $\mathbf{U}_j$,

$$p_j = \frac{1}{M+1}\left(1 + \sum_{m=1}^{M} \mathbb{1}(Z_j(\mathbf{X},\mathbf{Y}) \geq Z_j(\mathbf{U}_j^{(m)},\mathbf{Y}))\right).$$

Note that for the choice of $Z_j = \mathcal{L}$, the empirical risk of a model, this p-value computation is exactly the p-value computation from eq. (2). Recall our proof of uniform null p-values for CONTRA from prop. 1. If $Z_j(\mathbf{X},\mathbf{Y})$ and $Z_j(\mathbf{U}_j,\mathbf{Y})$ are equal in distribution, p_j will be uniform under the null. Thus, the closer W_j is to satisfying the flip-sign property, the closer null p-values are to being uniformly distributed.

If the empirical KL terms E_j are greater than 0, then the left hand side of eq. (5) will not be bounded, meaning W_j could have a positive or negative bias. As we discuss in the main text, when using the empirical risk of a model as Z_j, the bias of W_j tends to be negative, as the models exhibit higher losses on out-of-distribution data. A negative bias leads to deflated p-values, and ultimately inflated FDR.

Support mismatch between $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$ and $\hat{q}_{\text{cc}}^{(j)}$. When $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$ and $\hat{q}_{\text{cc}}^{(j)}$ have mismatched supports, E_j used in the Barber et al. (2020) proof above can be ∞ if $\hat{q}_{\text{cc}}^{(j)}$ puts no mass where $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$ puts non-zero mass. This means that the bound in eq. (5) can be vacuous and lead to highly non-uniform p-values. We will illustrate an example where HRTs realize this bound and yield low p-values even in the case of null covariates.

Consider the case of HRTs that use the most powerful test statistic, as shown by Katsevich and Ramdas (2020): the log-likelihood of $\hat{q}_{\text{model}}(\mathbf{y} \mid \mathbf{x})$. Let $\mathbf{x}_j$ be a null covariate. The model $\hat{q}_{\text{model}}$ is fit to the data $(\mathbf{X},\mathbf{Y})$ and its log-likelihood is well-defined; $\log \hat{q}_{\text{model}}(\mathbf{y} = \mathbf{y}^{(i)} \mid \mathbf{x} = \mathbf{x}^{(i)})$ will be in $(-\infty,0]$ if $\mathbf{x}^{(i)} \sim q(\mathbf{x})$ and $\mathbf{y}^{(i)} \sim q(\mathbf{y} \mid \mathbf{x} = \mathbf{x}^{(i)})$. Assume that there is high dependence between $\mathbf{x}_j$ and some non-null $\mathbf{x}_k$, so $\hat{q}_{\text{model}}$ exhibits spurious dependence on $\mathbf{x}_j$. If $\hat{q}_{\text{model}}$ is evaluated on a sample $(\widetilde{\mathbf{x}}_j^{(i)},\mathbf{x}_{-j}^{(i)},\mathbf{y}^{(i)})$ where $\widetilde{\mathbf{x}}_j^{(i)} \sim \hat{q}_{\text{cc}}^{(j)}$ is not in the support of $q(\mathbf{x}_j \mid \mathbf{x}_{-j} = \mathbf{x}_{-j}^{(i)})$, then $\hat{q}_{\text{model}}(\mathbf{y} = \mathbf{y}^{(i)} \mid \mathbf{x}_j = \widetilde{\mathbf{x}}_j^{(i)},\mathbf{x}_{-j} = \mathbf{x}_{-j}^{(i)})$ can be arbitrarily small. This means that $\mathcal{L}(\mathbf{U}_j^{(m)},\mathbf{Y})$ can be arbitrarily large if even a single sample $(\widetilde{\mathbf{x}}_j^{(i)},\mathbf{x}_{-j}^{(i)},\mathbf{y}^{(i)}) \in \mathbf{U}_j^{(m)}$ contains coordinate $\widetilde{\mathbf{x}}_j^{(i)}$ that is out of support for $q(\mathbf{x}_j \mid \mathbf{x}_{-j} = \mathbf{x}_{-j}^{(i)})$. The more likely it is to draw a sample from $\hat{q}_{\text{cc}}^{(j)}$ that results in such an $\mathcal{L}(\mathbf{U}_j^{(m)},\mathbf{Y})$, the more like the p-value will be 0. A realistic example of such cases is when $\hat{q}_{\text{cc}}^{(j)}$ captures only a single mode of $q(\mathbf{x}_j \mid \mathbf{x}_{-j})$. This leads to null features being selected, inflating the FDR.

Contrast this scenario with CONTRA instead of an HRT. Recall that the log-likelihood of $\hat{q}_{\text{mix}}^{(j)}$, which consists of an equal mixture of $\hat{q}_{\text{null}}^{(j)}$ and $\hat{q}_{\text{model}}$, is always well-defined. Then $\mathcal{L}(\mathbf{X},\mathbf{Y})$ will be no greater than

$$\sum_{i=1}^{n_{\text{test}}} \log \frac{1}{2} \hat{q}_{\text{model}}(\mathbf{y} = \mathbf{y}^{(i)} \mid \mathbf{x} = \mathbf{x}^{(i)}),$$

and $\mathcal{L}(\mathbf{U}_j^{(m)},\mathbf{Y})$ will be no less than

$$\sum_{i=1}^{n_{\text{test}}} \log \frac{1}{2} \hat{q}_{\text{null}}^{(j)}(\mathbf{y} = \mathbf{y}^{(i)} \mid \mathbf{x}_j = \widetilde{\mathbf{x}}_j^{(i)},\mathbf{x}_{-j} = \mathbf{x}_{-j}^{(i)}).$$

This is because even if $\hat{q}_{\text{model}}(\mathbf{y} = \mathbf{y}^{(i)} \mid \mathbf{x}_j = \widetilde{\mathbf{x}}_j^{(i)},\mathbf{x}_{-j} = \mathbf{x}_{-j}^{(i)}) = 0$, the term $\hat{q}_{\text{null}}^{(j)}(\mathbf{y} = \mathbf{y}^{(i)} \mid \mathbf{x}_j = \widetilde{\mathbf{x}}_j^{(i)},\mathbf{x}_{-j} = \mathbf{x}_{-j}^{(i)})$ will not be 0 as $\hat{q}_{\text{null}}^{(j)}$ is fit to samples from $\hat{q}_{\text{cc}}^{(j)}$. This means that the log-probability of the mixture will be in $(-\infty,0]$ and is therefore well-defined.

A consequence of the behavior of HRTs and CONTRA in these scenarios is that HRTs can arbitrarily deflate p-values. Contra alleviates this issue as it restricts the amount p-values can be deflated. As we show in our experiments, this leads to better FDR in practice.