Tim B.

The overlooked need for negative controls in conditional independence testing

1 Introduction

Conditional independence (CI) tests asses the association between two variables (e.g., a genetic variant and a phenotype) while controlling for a vector of confounders (e.g., population structure). CI tests are among the most fundamental and widely-used hypothesis tests in the sciences, technology, and other areas. Despite their importance, CI tests pose a basic and unavoidable difficulty: assumption-free CI testing is impossible (Shah and Peters 2020; Kim et al. 2021). Put differently, all valid CI tests must make an assumption (or set of assumptions) about the data-generating process. In practice these assumptions are seldom checked, greatly limiting the trustworthiness of results. This is not due to negligence; to the contrary, checking the assumptions of CI tests is fraught with difficulties, as we will show. We therefore face a dilemma as data analysts: we are obligated to check the assumptions of CI testing procedures to ensure reliableness of the results, but very often this is challenging (or impossible) to do.

Our core thesis is that "negative controls" – external samples for which the null hypothesis is known to be true, roughly – are crucially important (and in some cases *required*) for verifying the assumptions of CI testing procedures, enabling rigorous inference. We work in the contemporary "high-multiplicity" setting in which we seek to test thousands (or more) of hypotheses and produce a discovery set with guaranteed false discovery rate (FDR) control (Benjamini and Hochberg 1995; Li and E. J. Candès 2021).

First, we briefly summarize the vast and growing landscape of CI testing procedures, omitting from our discussion those procedures that do not enable the selection of critical regions (and thus the control of FDR). We argue that negative controls play (or ought to play) a crucial role in high-multiplicity CI testing. We describe two broad types of negative controls — "experimental" negative controls and "in silico" negative controls — and argue that, although the former are superior statistically, the latter can be constructed directly from the data in many applications and are therefore more broadly available.

Next, we introduce several new strategies for working effectively with negative control data in high-multiplicity hypothesis testing problems. (The

discussion here extends beyond CI testing.) We propose to calibrate the testing procedure against *both* the empirical negative control distribution *and* the theoretical null distribution, satisfying an appealing double-robustness property. We also introduce a simple and practical method for assessing whether a given procedure is robust to inflation (or deflation) in the tail of the empirical null distribution. Finally, we introduce the "symmetry plot" (or "s-plot"), a nonparametric analogue of the commonly-used quantile-quantile plot (qq-plot) to aid in the application of the above methods.

As an auxiliary contribution, we introduce a new class of fast and powerful Gaussian test statistics for use in a broad range of existing CI testing methods, including the conditional randomization test, the conditional permutation test, and the local permutation test (E. Candès et al. 2018; Berrett et al. 2020; Kim et al. 2021). The idea behind these statistics is to repeatedly fit OLS, ridge regression, or additive spline models to the permuted (or resampled) data via an online QR decomposition, achieving high speed and power. We apply the ideas that we introduce to analyze a new and important type of biological data that combines CRISPR genome editing with single-cell RNA sequencing.

2 A brief taxonomy of CI testing methods

There exist myriad methods for testing CI; see Kim et al. 2021 for a recent review. Inspired by Kim et al. 2021, we consider a simplified "taxonomy" of CI testing methods that consists of four categories: (i) standard parametric, (ii) model-X, (iii) matching, and (iv) nonparametric asymptotic. The first two categories are model-based (in that they assume a model — possibly up to a parametric family — for the data generating process), and the latter two are model-free. Figure 1 illustrates these categories and provides an example of each.

We briefly review CI testing methods that fall into these categories and highlight their underlying assumptions. Standard parametric approaches assume that the distribution $\mathcal{L}(Y|X, Z)$ of Y given X and Z is known up to parametric family; examples include GLMs. Inference (typically) is based on the CLT, which, in addition to correct model specification, requires the number of samples to be large relative to the number of parameters. Model-X approaches make the rather strong (and sometimes reasonable) assumption that the distribution $\mathcal{L}(X|Z)$ of X given Z is known exactly. An example is

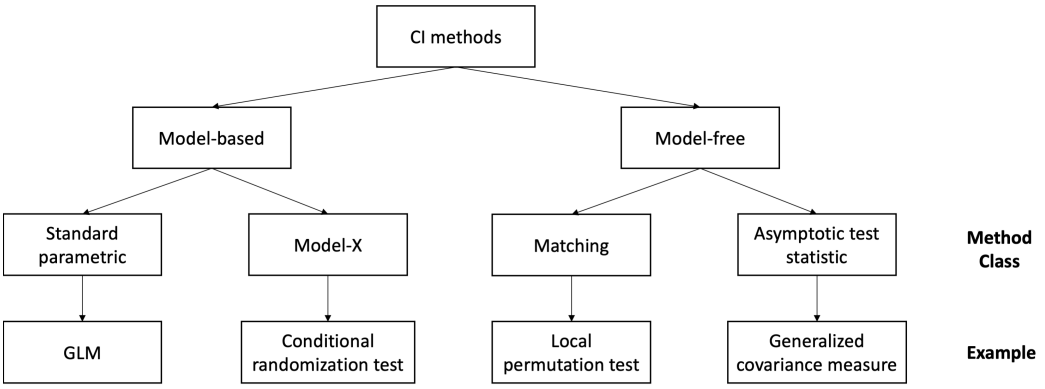

Figure 1: An abbreviated taxonomy of CI testing methods.

the conditional randomization test (CRT; E. Candès et al. (2018)). Inference in model-X methods is exact in finite samples.

Matching-based approaches test for CI by grouping observations based on their vector of confounders Z. An example is the recently-proposed local permutation test (LPT; Kim et al. (2021)). The LPT approximates the joint distribution (Y, X, Z) by $(Y, X, \tilde{Z})$, where $\tilde{Z}$ is a discretized version of Z. The LPT assumes that $X \perp\!\!\!\perp Y|\tilde{Z}$ implies $X \perp\!\!\!\perp Y|Z$ and uses a permutation-based approach to test the former hypothesis. Finally, the nonparametric asymptotic approach leverages nonparametric regression and/or density estimators to construct a test statistic whose asymptotic distribution is known under the null hypothesis. An example is the generalized covariance measure (GCM; Shah and Peters (2020)). GCM estimates the regression functions $f(z) := \mathbb{E}[X|Z = z]$ and $g(z) := \mathbb{E}[Y|Z = z]$ using (possibly black box) regression estimators and constructs an asymptotically Gaussian test statistic by taking a normalized product of the residuals. GCM assumes that the regression functions f and g can be estimated sufficiently fast.

We emphasize that each of the above approaches makes an assumption or set of assumptions about the joint distribution $\mathcal{L}(X, Y, Z)$. These assumptions are mathematically concrete and could in theory be checked by an oracle with knowledge of the joint distribution. In practice, of course, we do not have knowledge of (X, Y, Z) and thus must consider other strategies for assumption checking.

3 The crucial importance — and difficulty — of assumption checking in CI testing

Assumption checking is of crucial importance in CI testing. Consider the local permutation test (LPT) of Kim et al. 2021 as an illustrative example. [Insert a small simulation study applying LPT to synthetic data for which the assumptions of LPT are violated. Show that, in the absence of assumption checking, type-I error inflation and FDR violation can be severe.]

Unfortunately, the assumptions underlying LPT are uncheckable (absent additional, external information). Let $P^n_{X,Y,Z}$ denote the product distribution of the observed data $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$. Next, let $Q^n_{X,Y,\tilde{Z}}$ denote the distribution that results from discretizing Z. Finally, let $\tilde{Q}^n_{X,Y,\tilde{Z}}$ denote the closest distribution (in TV distance) to $Q^n_{X,Y,\tilde{Z}}$ in the space of distributions for which $X \perp\!\!\!\perp Y | \tilde{Z}$. Kim et al. (2021) show that the excess type-I error of LPT is bounded above by $d_{\text{TV}}\left(\tilde{Q}^n_{X,Y,\tilde{Z}}, \tilde{Q}^n_{X,Y,\tilde{Z}}\right)$. They furthermore derive an explicit, analytic upper bound that holds if the conditional distributions $\mathcal{L}(X|Z=z)$ and $\mathcal{L}(Y|Z=z)$ are sufficiently smooth as z varies. Although theoretically informative, these bounds cannot be evaluated in practice, as the densities $\mathcal{L}(X,Y,Z)$, $\mathcal{L}(X|Z)$, and $\mathcal{L}(Y|Z)$ are unknown. We highlight LPT only to provide the reader with a concrete example; other model-free approaches (e.g., GCM) likewise place assumptions on $\mathcal{L}(X,Y,Z)$ that cannot be checked in practice.

Assumption checking in model-based approaches likewise is fraught with challenges, especially in high dimensions. Suppose that we test CI by fitting a linear model to the data. Goodness-of-fit checks that work well in low dimensions — e.g., calculating likelihood-based metrics such as AIC, estimating predictive performance via cross-validation, etc. — break down in high-dimensions. For example, overfitting can render meaningless metrics based on likelihood, and predictive performance does not necessarily provide much information about estimation performance. (In fact, prediction and estimation risk are inversely related in the high-dimensional linear model; Dobriban and Wager (2018).) Therefore, model-based approaches pose similar challenges with respect to model checking to model-free approaches, especially in high-dimensions. [Can we prove a general result about the impossibility of checking assumptions of CI tests?]

4 The need for negative controls

We make two distinctions. First, we distinguish between "calibration data" and "test data." Calibration data are negative control data on which we assess calibration, and test data are data on which apply our procedure with the objective of obtaining FDR control. Ideally, the calibration and test data are different, but in some applications they may coincide.

- single-cell CRISPR screens
  - Experimental controls typically are available.
  - When they are not, pair perturbations to genes on different chromosomes, avoiding transcription factors.
- eQTLs
  - Pair SNPs to genes on different chromosomes, again avoiding transcription factors.
- multimodal single-cell data (e.g., chromatin accessibility and gene expression)
  - Pair chromatin regions with genes on other chromosomes, again avoiding transcription factors.
- bulk RNA-seq
  - Really, we are testing a conditional independence hypothesis here: is there differential expression *conditional* on the library size? Most genes are not differentially expressed, and so they can serve as negative controls. Here, calibration and test data coincide.

5 Strategies for working effectively with negative controls

We hope that the reader is convinced that (i) negative controls are of crucial importance in conditional independence testing and that (ii) negative controls — either experimental or *in silico* — are available or can be constructed for many important datasets, especially those in genomics and genetics. A

natural question, then, is "How can we work effectively with negative control data?" We take steps toward answering this question. We propose to calibrate the test procedure against *both* the empirical negative control distribution *and* the theoretical null distribution, satisfying a double-robustness property. First, we take a slight detour to introduce "symmetry plots," nonparametric analogues of qq-plots.

5.1 Symmetry plots (s-plots)

Symmetry plots ("s-plots") are visual diagnostics that track the extent to which an empirical distribution is symmetric about zero. Let $X_1, \ldots, X_B \sim F$ be samples from some common distribution with density function f. Assume that B is even (otherwise, throw out the median). Let $X_{(1)}, \ldots, X_{(B)}$ be the order statistics of $X_1, \ldots, X_B$. Consider the set

$$\mathcal{S} := \{(-X_{(1)}, X_{(B)}), \ldots, (-X_{(i)}, X_{(B-i+1)}), \ldots, (-X_{(B/2)}, X_{(B/2+1)})\},$$

i.e., pair the negative of the first order statistic with the last order statistic, the negative of the second order statistic with the penultimate order statistic, etc. An s-plot is a plot of the set $\mathcal{S}$. If the underlying density function f is symmetric about zero, then the points fall approximately onto the identity line $y = x$. Figure 2 shows a histogram, s-plot, and qq-plot of $B = 10,000$ draws from a double-exponential distribution. Clearly, the histogram (left) is not Gaussian (blue curve indicates Gaussian MLE); accordingly, the qq-plot (right) exhibits inflation in the tail. However, the double-exponential distribution is symmetric about zero; hence, the s-plot closely follows the identity line.

The rationale for the s-plot is as follows. The expected value of $X_{(i)}$ should be a quantile of order $(i - 1/2)/B$. Meanwhile, the expected value of $X_{(B-i+1)}$ should be a quantile of order $[B - i + 1/2]/B$. Note that

$$[i - 1/2]/B + [B - i + 1/2]/B = 1.$$

Because f is symmetric about the origin, it follows that

$$-F^{-1}([i - 1/2]/B) = F^{-1}([B - i + 1/2]/B).$$

Thus,

$$\mathbb{E}[-X_{(i)}] \approx -F^{-1}([i - 1/2]/B) = F^{-1}([B - i + 1/2]/B) \approx \mathbb{E}[X_{(B-i+1)}],$$

implying the points in the s-plot lie approximately along the identity line.

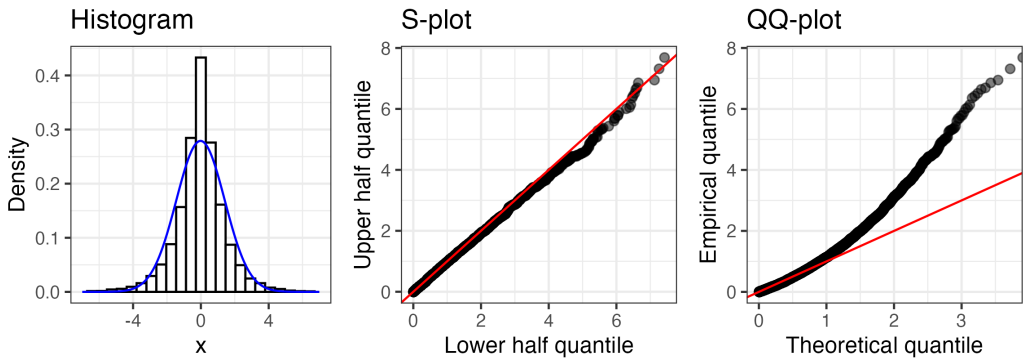

Figure 2: Histogram, s-plot, and qq-plot (only right tail shown) of $B = 10,000$ draws from a double-exponential distribution.

5.2 The Barber-Candés method and symmetry-preserving transformations

The Barber-Candés method is a flexible method for FDR control.

References

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society* 57.1, pp. 289–300.

Berrett, Thomas B. et al. (2020). "The conditional permutation test for independence while controlling for confounders". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 82.1, pp. 175–197.

Candès, Emmanuel et al. (2018). "Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 80.3, pp. 551–577.

Dobriban, Edgar and Stefan Wager (2018). "High-dimensional asymptotics of prediction: Ridge regression and classification". In: *Annals of Statistics* 46.1, pp. 247–279.

Kim, Ilmun et al. (2021). "Local permutation tests for conditional independence". In:

Li, Shuangning and Emmanuel J. Candès (2021). "Deploying the Conditional Randomization Test in High Multiplicity Problems". In: pp. 1–43.

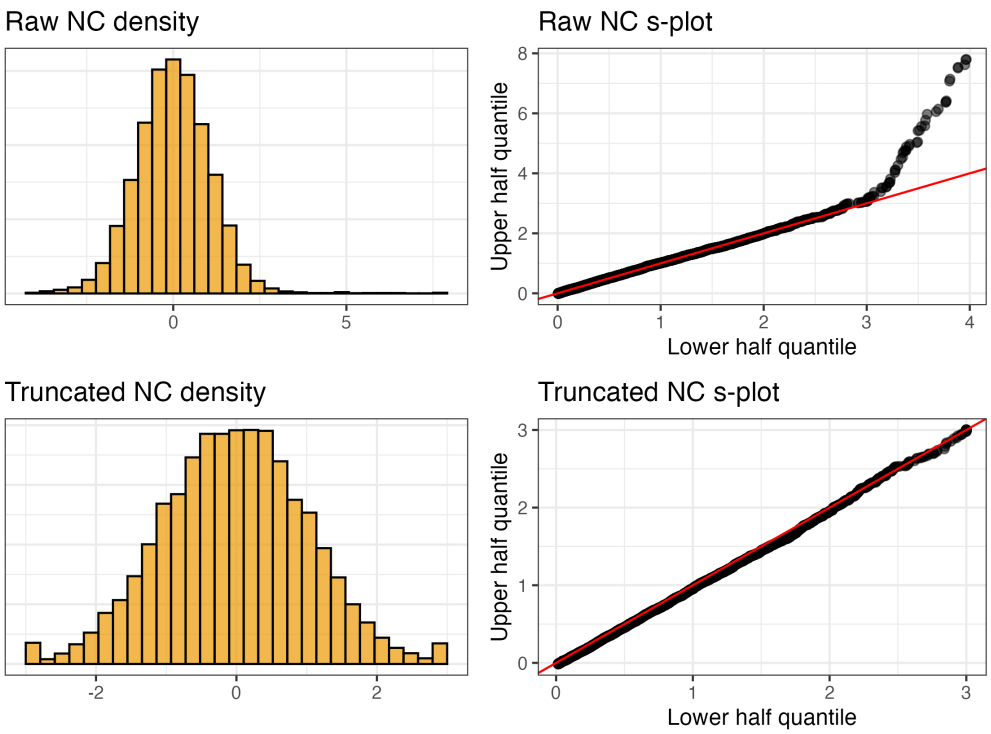
Raw NC density
Raw NC s-plot
Upper half quantile
Lower half quantile
Truncated NC density
Truncated NC s-plot
Upper half quantile
Lower half quantile

Figure 3:

Shah, Rajen D. and Jonas Peters (2020). "The hardness of conditional independence testing and the generalised covariance measure". In: *Annals of Statistics* 48.3, pp. 1514–1538.