Tim B.

The (overlooked?) importance of negative controls in conditional independence testing

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 task is challenging (or even impossible) to carry out.

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 misspecification in the tail of the empirical null distribution. 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 algorithm, achieving high power and speed. Finally, we illustrate the ideas in this work by applying them to analyze a new kind of biological data that combines CRISPR genome editing with single-cell RNA sequencing.

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.

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.

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.