Machine hypothesis testing in multi-axis, high-throughput data

Tim B

February 6, 2022

1 The explosion of multi-axis high-throughput data

A transformation occurred in statistics at the turn of the twentieth century. New biotechnologies, including as microarrays, GWAS, and bulk RNA-seq, yielded tens of thousands (or more) of measurements, necessitating the development of new statistical approaches to handling many hypotheses (e.g., FDR; Figure 1). These technologies yielded many measurements along a *single* axis (Figure 2b-c). For example, in microarrays and bulk RNA-seq, expression level is measured for many genes ($Y_1, \ldots, Y_p$) across an experimental condition (e.g., diseased or healthy). In this sense microarray and bulk RNA-seq data are "wide." By contrast, in GWAS, many genetic variants ($X_1, \ldots, X_d$) are measured, their their association with a phenotype is tested. In this sense GWAS data are "tall." (Note that "tall" and "wide" here refer to the number of features measured along different axes of the data, not the dimension of the data. Whether a dataset is "tall" or "wide" in this context is orthogonal to whether it is low- or high-dimensional.) We note that scientists sometimes conduct multi-trait GWAS studies, but these studies are "low throughput" in the number of traits (i.e., a few traits, rather than tens or hundreds of thousands of traits, are examined).

Newer technologies, including eQTLs, multimodal single-cell assays, and single-cell CRISPR screens, make tens of thousands (or more) of measurements along *multiple* axes. For example, eQTLs measure tens of thousands of genes alongside hundreds of thousands of genetic variants. Multimodal

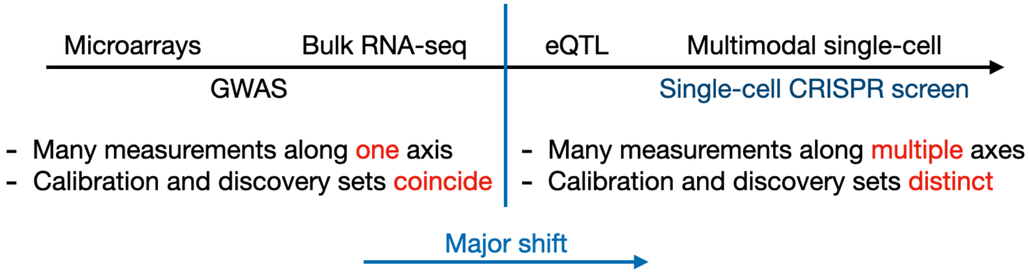
Microarrays
Bulk RNA-seq
GWAS
eQTL
Multimodal single-cell
Single-cell CRISPR screen
- Many measurements along one axis
- Calibration and discovery sets coincide
- Many measurements along multiple axes
- Calibration and discovery sets distinct
Major shift

Figure 1: Older technologies (e.g., microarrays, GWAS, and bulk RNA-seq) produce many measurements along a single axis. New technologies (e.g., eQTLs, multimodal single-cell experiments, and single-cell CRISPR screens) produce many measurements along *multiple* axes, changing how we can think about calibration. Single-cell CRISPR screens (blue) are an especially important emerging biotechnology that typically come equipped with *experimental* negative and positive controls.

single-cell assays measure chromatin accessibility across tens of thousands of locations alongside gene expressions (for example). And single-cell CRISPR screens measure thousands of CRISPR perturbations alongside gene expressions. These newer technologies yield datasets that are both "tall" and "wide" (Figure 2a). We call such "tall" and "wide" data "multi-axis high-throughput data."

We argue that the shift from single-axis to multi-axis high-throughput data is a significant one. In particular, the shift opens the door to new ways of thinking about calibration. On single-axis high-throughput data, the "calibration" and "discovery" sets coincide: we must use the same data to test for associations and assess calibration. Traditionally, statisticians have relied on sparsity assumptions for this purpose. By contrast, on multi-axis high-throughput data, we can construct distinct calibration and discovery sets. This enables us to take a prediction-inspired approach to testing. Single-cell CRISPR screens, a new kind of genomic data rapidly gaining importance, typically come equipped with *experimental* negative *and* positive controls.

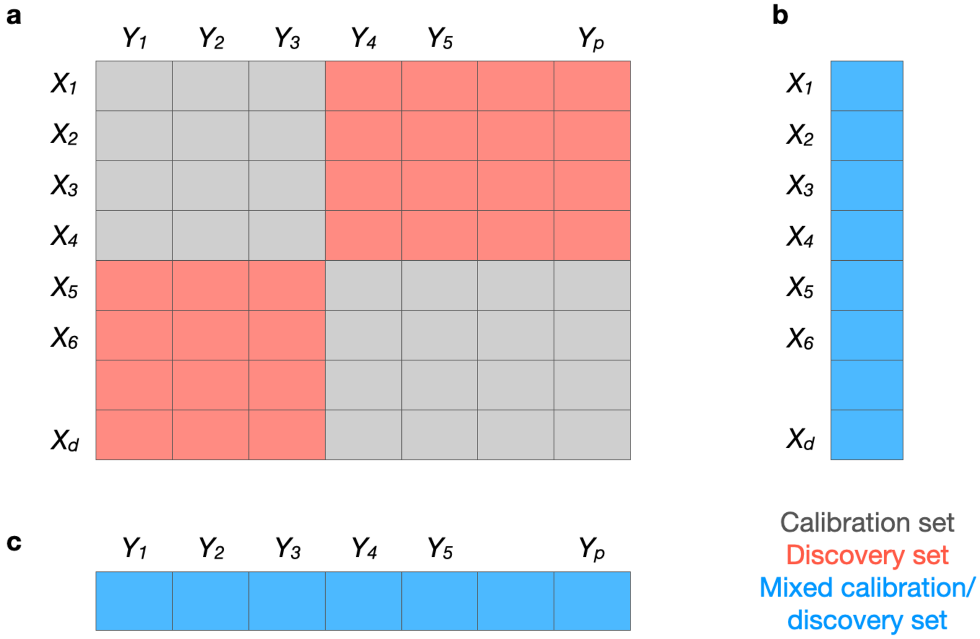

Figure 2: Comparison of three different types of data: (a) tall and wide, (b) tall, and (c) wide. Examples of (a) include eQTL, multimodal single-cell, and single-cell CRISPR screen data; an example of (b) is GWAS data; and examples of (c) include microarray and bulk RNA-seq data.

2 A framework for negative controls

Our goal is to formalize and unify the use of negative controls in conditional independence testing. Here, we consider the simplest, nontrivial case. **Background**. As a review, suppose we seek to test the hypothesis $X \perp\!\!\!\perp Y | Z$ using data $\{(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)\}$. Let $f_{X,Y,Z}$ be the joint density of (X_i, Y_i, Z_i). The null hypothesis is true if and only if

$$f_{X,Y|Z} = f_{Y|Z} f_{X|Z}$$

for some densities $f_{Y|Z}$, and $f_{X|Z}$. The null hypothesis, of course, is composite. Let $f^{\text{null}}_{X,Y|Z}$ be the set of densities that factors in the above manner. Next, let $\phi : (\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d)^n \to \{0, 1\}$ be a test of the null hypothesis. We say that ϕ is level α if

$$\sup_{f \in f^{\text{null}}_{X,Y|Z}} \mathbb{E}_f \left[\phi \left(\{(X_i, Y_i, Z_i)\}_{i=1}^n \right) \right] \leq \alpha.$$

Statistical framework. We now put ourselves in the "high-multiplicity" setting. Throughout, we condition on Z. Let $\{f^1_{X,Y|Z}, \dots f^p_{X,Y|Z}\}$ be p conditional densities. Let $\{D_1, \dots, D_p\}$ be p independent (conditional on Z) datasets, where

$$D_i = \{(X^i_1, Y^i_1, Z_1), \dots, (X^i_n, Y^i_n, Z_n)\},$$

and

$$(X^i_1, Y^i_1), \dots, (X^i_n, Y^i_n) \sim f^i_{X,Y|Z}.$$

In other words, $D_i \in \mathbb{R}^{n \times (2+d)}$ is a dataset consisting of the columns $X^i := (X^i_1, \dots, X^i_n)$, $Y^i := (Y^i_1, \dots, Y^i_n)$, and $Z := (Z_1, \dots, Z_n)$; the rows $\{(X^i_j, Y^i_j)\}_{j=1}^n$ of the first two columns of D_i are drawn i.i.d. from the conditional distribution $f^i_{X,Y|Z}$; and the final column $Z = (Z_1, \dots, Z_n)$ of D_i is held constant across datasets.

For $i \in \{1, \dots, p\}$, let H_i be the hypothesis that

$$f^i_{X,Y|Z} = f^i_{X|Z} f^i_{Y|Z}$$

for some $f^i_{X|Z}$ and $f^i_{Y|Z}$. In other words, H_i is the hypothesis that the conditional independence property holds for the ith density. We test hypothesis H_i by applying a conditional independence test to dataset D_i. Let $\mathcal{N} \subset \{1, \dots, p\}$ be the set of hypotheses for which the null hypothesis is true, and let $|\mathcal{N}|$ denote the number of true null hypotheses.

In many applications the conditional densities $f^i_{X|Z}$ and $f^i_{Y|Z}$ are similar across the is. For example, in a single-cell CRISPR screen experiment, $\{f^i_{Y|Z}\}_{i=1}^p$ are the gene expression densities (conditional on the technical factors). These densities are similar across features: aside from some gene-to-gene variability in mean expression level and dispersion, the gene expression densities are all sparse, discrete counts.

We can encode this similarity across features using a Bayesian approach. Let $\{f_{X|Z}(\theta_X) : \theta_X \in \mathbb{R}^{m_x}\}$ be a family of densities parameterized by $\theta_X \in R^{m_x}$. Similarly, let $\{f_{Y|Z}(\theta_Y) : \theta_Y \in \mathbb{R}^{m_y}\}$ be a family of densities parameterized by $\theta_Y \in \mathbb{R}^{m_y}$. Next, let G_x and G_y be probability distributions. We assume the following hierarchical model:

$$\begin{cases} \theta_X^1, \ldots, \theta_X^p \sim G_x \\ f_{X|Z}^1 = f_{X|Z}(\theta_X^1), \ldots, f_{X|Z}^p = f_{X|Z}(\theta_X^p) \\ \theta_Y^1, \ldots, \theta_Y^p \sim G_y \\ f_{Y|Z}^1 = f_{Y|Z}(\theta_Y^1), \ldots, f_{Y|Z}^p = f_{Y|Z}(\theta_Y^p). \end{cases}$$

Hence, for $i \in \mathcal{N}$, the density factors as

$$f_{X,Y|Z}^i = f_{X|Z}^i f_{Y|Z}^i = f_{X|Z}(\theta_X^i) f_{Y|Z}(\theta_Y^i).$$

Negative controls. Negative controls exploit the similarity across features to assess calibration. For $i \in \{1, \ldots, p_{\text{nc}}\}$, let $\tilde{f}_{X,Y|Z}^i$ denote the density of the ith negative control. By definition, the null hypothesis holds true for the negative controls, and so

$$\tilde{f}_{X,Y|Z}^i = \tilde{f}_{X|Z}^i \tilde{f}_{Y|Z}^i$$

for some densities $\tilde{f}_{X|Z}^i$ and $\tilde{f}_{Y|Z}^i$. Let $\tilde{D}_i \in \mathbb{R}^{2+d}$ denote the dataset generated by $\tilde{f}_{X,Y|Z}^i$, i.e.

$$\tilde{D}_i = \{(\tilde{X}_1^i, \tilde{Y}_1^i, Z_1), \ldots, (\tilde{X}_n^i, \tilde{Y}_n^i, Z_n)\},$$

where

$$(\tilde{X}_1^i, \tilde{Y}_1^i), \ldots, (\tilde{X}_n^i, \tilde{Y}_n^i) \sim \tilde{f}_{X,Y|Z}^i,$$

and Z is the same as before. Our core assumption is that the negative control datasets are "similar" to the null datasets, which we formalize below.

Key assumption: *Let $\tilde{\theta}_X^1, \ldots, \tilde{\theta}_X^{p_{nc}} \sim G_x$ and $\tilde{\theta}_Y^1, \ldots, \tilde{\theta}_Y^{p_{nc}} \sim G_y$. We assume that that $\tilde{f}_{X|Z}^i = f_{X|Z}(\tilde{\theta}_X^i)$ and $\tilde{f}_{Y|Z}^i = f_{X|Z}(\tilde{\theta}_Y^i)$.*

In other words, we assume that the conditional densities $f_{X|Z}$ and $f_{Y|Z}$ are drawn i.i.d. from the *same* underlying distribution on both the negative control data *and* the null data. A key corollary results from the assumption.

Key corollary: *The null data $\{D_i : i \in \mathcal{N}\}$ and the negative control data $\{\tilde{D}_i\}_{i=1}^{p_{nc}}$ are identically distributed.*

This corollary implies that we can use the negative control data to assess calibration of the method on the rest of the data. To flesh this idea out, let $\phi_\alpha : (\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d)^n \to \{0, 1\}$ be a level α test of the null hypothesis. If $\mathbb{E}[\phi_\alpha(\tilde{D}_i)] \leq \alpha$ (i.e., ϕ_α controls type-I error at level α on the negative control data), it follows that $\mathbb{E}[\phi_\alpha(D_i)]$ for $i \in \mathcal{N}$ (i.e., ϕ_α controls type-I error on the null data). This is because $\tilde{D}_j \overset{d}{=} D_i$ for $i \in \mathcal{N}$. Pulling back, in using negative controls, we implicitly make the assumption that the negative control datasets have a similar (or identical) distribution to the null datasets. The framework presented here is one way of formalizing that assumption.

3 An analogy between prediction and CI testing

We propose a framework for CI testing with negative controls that is directly analogous to prediction. We propose the following steps:

1. Let $\mathcal{S} = \{D_1, \ldots, D_B\}$ be the B (possibly dependent) negative control datasets.

2. Partition the negative control datasets into train and validation sets in such a way that the two sets are independent (given Z). In other words, define sets $\mathcal{T}$ and $\mathcal{V}$ such that

   i. $\mathcal{T} \cap \mathcal{V} = \emptyset$.

   ii. $\mathcal{T} \cup \mathcal{V} = \mathcal{S}$.

   iii. $\mathcal{T} \perp\!\!\!\perp \mathcal{V}$.

	Prediction	CI testing
Basic unit	An example (X_i, Y_i, Z_i)	A negative control dataset $D_{ij} = \{(X_l^i, Y_l^j, Z_l)\}_{l=1}^n$
Algorithm	A CI test ϕ	A prediction function f
Ground truth	Label Y_i available	Dataset D_{ij} a negative control
Calibration set	Subset of (known ground truth) data on which we train predictor	Subset of (known ground truth) data on which we calibrate CI testing method
Validation set	Subset of (known ground truth) data on which we assess predictor performance	Subset of (known ground truth) data on which we assess CI method performance
Discovery set	Unlabeled data $\{(X_i, Z_i)\}$ on which we seek prediction	New data $\{(X_l^i, Y_l^j, Z_l)\}_{l=1}^n$ with unknown ground truth
Target metric	Type-I error as a function of α, $E(\alpha)$: $$E(\alpha) = \mathbb{P}(\phi(D_i) \leq \alpha)$$	Prediction risk R given metric d: $$R = \mathbb{E}[d(Y_i, f(X_i, Z_i))]$$
Empirical metric	Empirical type-I error	Empirical risk

4 Positive controls and partially synthetic positive controls

We can use positive controls (real or semi-synthetic) to boost power in the calibration phase.

5 Borrowing ideas from prediction

The analogy enables us to borrow potentially useful ideas from prediction:

- Robustness to distribution shift
- Hyperparameter tuning
- Cross validation