Deep IV with Whole-Omics Data to Identify Novel Carcinogenesis-Mediating Pathways

Project Milestone
Category: Health / Genomics

Jack Andraka
jandraka@stanford.edu

Billy Ferguson
billyf@stanford.edu

Charlie Walker
cwalker4@stanford.edu

1 Introduction

Breast cancer accounts for over 25% of cancer diagnoses and 15% of cancer-related deaths in women.[1] Ten percent of women with breast cancer have a family history of disease. Women with one premenopausal first-degree relative with breast cancer are at 3.3-fold greater risk than women without a family history,[2] demonstrating that there is a significant genetic contribution to breast cancer risk. To identify genetic factors associated with breast cancer, early studies employed linkage analysis and positional cloning in families with a familial history of breast cancer to discover highly penetrant susceptibility genes such as BRCA1&2.[3] Although these initial studies were successful and could explain about 20% of the familial risk of breast cancer,[4] they provided little insight into the role of genetics in nonfamilial breast cancer.

More recently, genome-wide association studies (GWAS) have identified over 80 loci significantly associated with sporadic breast cancer. However, these variants collectively only explain 16% of breast cancer heritability.[5] The inability of GWAS to identify a greater proportion of the genetic risk stems from many factors, including genotyping platform limitations in interrogating rare variation (primarily due to the running of several thousand to several million t-tests simultaneously). This is in contrast with the current hypothesis of genetic architecture that posits many gene variants acting in tandem to produce a phenotypic trait, each with small effect size.[6] The GWAS study design is also plagued by confoundedness, despite the development of many sophisticated statistical techniques to deal with complicated interactions. Simple prediction and correlation setups in breast cancer studies fail to account for this confoundedness (gene expression results in cancerous growth and cancerous growth impacts gene expression).

Instrumental variables (IV) have proven to be an adept statistical method for addressing these issues of confoundedness. As shown in Figure 1, IV uses an exogenous instrument z to identify direct causal relationships between some policy variable p on outcome y (a relationship confounded by latent effects e). In the case of genetics studies, Mendelian randomization of genetic variants offer up a promising instrument for estimating causal effects of gene expression on cellular phenotype, in this case if the cell is cancerous or not.[7] In this manner, it is possible to identify low-signal, rare variants that would typically not appear in GWAS analyses, thus providing a more comprehensive mapping of the transciptome of a cell to carcinogenesis.

Traditional IV experimental designs suffers from the limitation that they require a strong prior understanding of the data generating process (DGP), and are limited in accounting for complex interactions between covariates. Neural networks offer up a solution to this limitation, due to their ability to uncover complicated interactions between genes that are both near and distal.[8] This project provides a proof-of-concept implementation of the Deep-IV framework, applied to characterizing the causal effect of gene expression on carcinogenesis in breast cancer using random, simple nucleotide variants (SNVs).

2 Algorithm

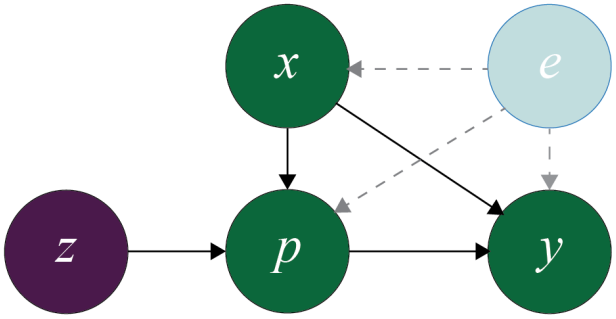
2.1 Overview

We implement the method described in "Deep IV: A Flexible Approach for Counterfactual Prediction" to identify the effect of a policy variable p on outcome y under confoundedness.[5] Instrumental variables (IV) are a well-developed tool for remedying endogeneity, but require a strong prior understanding of the data generating process and are not well equipped to deal with a large number of covariates. Deep IV promises to marry the best qualities of DNNs and IV, and we believe GWAS are a prime use case.

To perform IV analysis one needs to find an exogenous variable that affects the outcome variable only through the endogenous covariate of interest. More specifically, the instrument z must be conditionally independent of the error (Figure 1).

Traditional IV can be estimated through a procedure

Figure 1: Generalized Deep IV



called two-stage least squares (2SLS): in the first stage you regress your endogenous variable of interest, p, on the exogenous instrument, z, to create a predicted $\hat{p}$ constructed only with the exogenous variation of z. In the second stage, you regress your outcome variable, y, on the predicted $\hat{p}$ from the first stage.[1]

2.2 Model Architectures

To adapt this procedure for the DeepIV method, we replace the two-stage least squares with two-stage multilayer perceptrons. We call the first stage our *policy network* and the second our *response network*. Our first iteration is modeled off of the specification in the appendix of the seminal DeepIV paper from Hartford et al. (2017).[6]

The policy and response networks in our simulations below have three hidden layers with 128, 64, and 32 hidden units respectively. The policy network takes in the exogenous z and control variables x as input to predict p using tanh acitvation functions for each layer and a Mixture of Gaussian output with 10 components. The predicted $\hat{p}$ from the policy network and x are then fed into the response network to yield predictions $\hat{y}$. The repsone network performs ReLU activation functions for the three hidden layers and linear activation for the output.

Both networks use Adam Optimization with learning rate $= 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e-08$. For training we use L2 weight decay with penalty parameter 0.001, a dropout rate of $\min(1000/(1000+n); 0.5)$, $(1.5 \times 10^6/n)$ epochs, and batch size 100.

[1] Chapter 4 of Angrist and Pischke's *Mostly Harmless Econometrics: an Empiricist's Companion* provides a good introduction to instrumental variables.

3 Experiment

3.1 Motivation and Design

As a proof-of-concept, we evaluate our approach on simulated data, testing DeepIV's ability to recover an underlying causal relationship in a low-dimensional domain. We compared our DeepIV architecture to a single-stage feed-forward network with the same architecture as our DeepIV *response* stage.

3.2 Simulated Data

Our simulation models a DGP similar to that described in Section 2. For ease of explanation, we ground our simulation in a typical real-world IV application: estimating the effect of price on sales of some product (say, hotel bookings). We begin by assuming seven customer types, $s \in \{1, ..., 7\}$ which each exhibit different levels of price sensitivity. Customer price sensitivity varies according to a complex non-linear function of time.

$$\psi_t = 2\big((t-5)^4/600 + \exp\big[-4(t-5)^2\big] + t/10 - 2\big)$$
$$t \sim \text{unif}(0, 10)$$

Prices are a function of observed variable t and some instrument z, on the basis that the hotel chooses their price to move with average price sensitivity. In the hotel example, the high demand resulting from some unobserved confounding variable (e.g. a nearby conference) breaks the conditional independence between our policy variable p (price) and the latent effects e. We model this by generating our errors e with parameter ρ that denotes the correlation between p and e; in other words, it reflects the extent of endogeneity in our population model. Our outcome variable y (sales) is then generated as:

$$y = 100 + (100 + p)s\psi_t - 2p + e$$
$$p = 25 + (z+3)\psi_t + v$$
$$z, v \sim \mathcal{N}(0, 1)$$
$$e \sim \mathcal{N}(\rho v, 1 - \rho^2)$$

The causal relationship we wish to uncover is $h(t, s, p) = (10 + p)s\psi_t - 2p$, but the correlation between the error e and p in our population model violates the unconfoundedness assumption necessary for causal interpretation.

Since we have our ground truth causal relationship, we evaluate our model by first generating features $[t, s, p]$, but then change p to a fixed grid of price values p'. This allows us to compare our predicted sales, $\hat{h}$ against the ground truth h.

3.3 Results

Figures 2 and 3 summarize the results of our simulations for two different n, giving out-of-sample MSE as we vary

Figure 2: Causal Performance (n=1000)

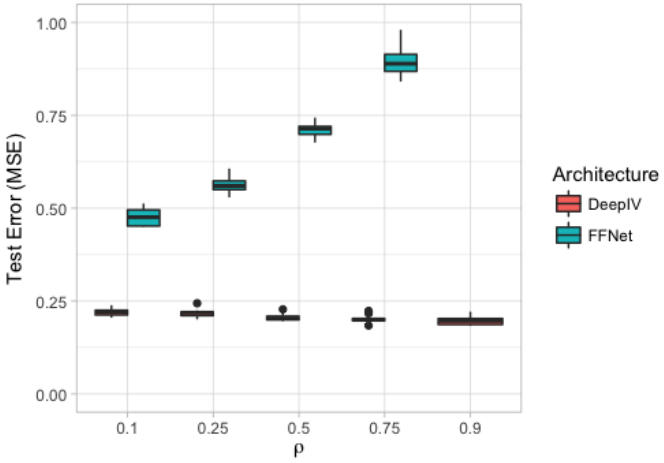
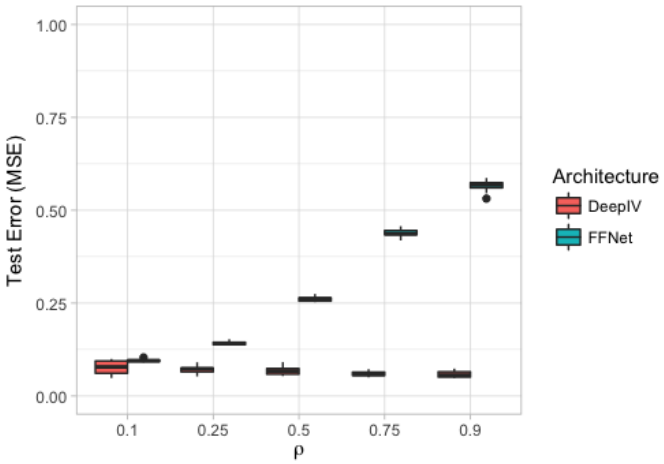


Figure 3: Causal Performance (n=10,000)



the level of endogeneity in our DGP. Each model was fit on 10 random samples from the DGP for each ρ and sample size. The performance of the DeepIV architecture was largely unaffected by the increase in ρ, reflecting the resilience of this architecture to confounding latent variables. On the other hand, the Feed-Forward Network did a poor job of recovering the true counterfactual function we were testing with this simulation.[2]

4 Data

For this project we are employing whole-genome and whole-transcriptome data from the Genotype-Tissue Expression (GTEx) project and The Cancer Genome Atlas (TCGA). The GTEx dataset includes 10,294 samples that are non-cancerous with matched whole-genome and whole-transcriptome data. The TCGA dataset includes 3,681 breast cancer samples with matched whole-genome and whole transcriptome data. For both of these datasets there is a total of 22,147 genes measured with 3,142,246 mutations recorded between the two datasets.

The transcriptome dataset for both GTEx and TCGA quantified 60,239 mRNAs, measured in transcripts per kilobase million. Our experimental setup will only use variants in transcription factors (1,391 genes) with a total of 6,495 mutations (ranges from 0-29 mutations per transcription factor). We will also employ L1 regularization for the second stage of our Deep-IV framework to eliminate highly correlated genes (reduce effects of linkage disequilibrium) and reduce our feature space to prevent overfitting.

5 Next Steps

This milestone was a proof of concept, showing that the DeepIV method effectively uncovers causal relationships that are glossed over by a naive Feed Forward Neural Net. Crucially, first implementing our architecture on simulated data provided a sanity check of our code before moving forward with our real life dataset where the true causal relationship is unknown.

Just recently we were granted access to breast cancer genetic data. The path forward entails cleaning this data, running our baseline model described above on the genetic data, and then tuning our hyperparameters. Additionally, we would like to include the standard 2SLS as another baseline to judge the effectiveness of the DeepIV architecture against the standard IV method.

[2]Note that this model would have performed well if we had been testing its ability to estimate $h(t, s, p) + \mathbb{E}[e|p]$, i.e. if we were not trying to make counterfactual predictions.