Dissertation

By

XIAOLIU WU
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

James Sharpnack

Committee Member 1

Thomas Lee

Committee Member 2

Krishnakumar Balasubramanian

Committee Member 3

Committee in Charge

2022

To my parents Jinzhi Liu and Yue Wu, my grandparents Yulan Zhao and Shouyi Liu, and my partner Yuanyuan Li, without your support and love, I would not be here.

Contents

Abstract

This dissertation is a combination of two bodies of work in modern statistical inference. The first work introduces and studies a novel conditional Independence test using neural networks. The second work is applied data analyses for the Healthy Davis Together (HDT) Program, for wastewater testing for Covid-19 infections and an analysis of the impact of the college reopening. We propose a neural Ising model to test conditional independence between binary random variables X and Y given a potentially complex random variable Z such as text or images. The method uses the score test statistic and employs a computationally efficient score-based bootstrap procedure [**20**] to generate the p-value. We extend the method to the multi-class X and Y by replacing the Ising model with a restricted Boltzmann machine. Empirical studies show that our model has high power against H_1 and reliable type-I error control on both simulated and real-world data. We derive the asymptotic separability of the score-test statistics under the Ising model.

On the applied side, we first summarize our collaboration with the wastewater team at Healthy Davis Together (HDT) initiative working on the wastewater monitoring project [**36**]. We provide a Bayesian Ct value imputation method via the EM-MCMC algorithm wrapped in a user-friendly API. The algorithm is able to produce Ct values matching the overall trend of the clinical data and has a stronger correlation with the clinical data when compared with existing methods [**36**]. The other data analysis project at HDT is measuring the impact of college reopening on the COVID-19 outbreak level in their home county. The coronavirus disease 2019 (COVID-19) pandemic has dramatically impacted the 2020-2021 academic year in universities across the country, and conversely, college reopening has disrupted the course of the pandemic. We investigated COVID-19 hotspot events in "college counties" which we defined as counties with at least 10% of its population composed of undergraduate students. We found that increments in cases could not be attributed to random chance by performing multiple hypothesis testing. Increments in confirmed cases among college counties from mid-August to mid-September were significantly higher than comparable non-college counties. After this period of reopening, hotspots of confirmed cases did not differ between counties, despite the college-town designation. Class setting (i.e., In-Person, Hybrid, Online) seemed to be associated with hotspot activity. We found no evidence to support an association between testing efforts and hotspots.

Acknowledgments

I'd like to thank my advisor James Sharpnack for his patience and guidance throughout my Ph.D. journey. I am also grateful to Professor Wolfgang Polonik for leading me into the field of statistics. Thank everyone who helps me to achieve what I have accomplished thus far.

CHAPTER 1

Introduction

The dissertation covers three projects which cover both statistics methodology and applied data analysis. Here we provide a brief background of each project. For detailed motivations, literature reviews, and setup, refer to the corresponding chapter.

Chapter 2 covers the development of the novel neural Ising model / Boltzmann machine for testing conditional independence of categorical X and Y given Z. Formally, the problem we study is

$$H_0 : X \perp\!\!\!\perp Y|Z \text{ versus } H_1 : X \not\perp\!\!\!\perp Y|Z.$$

Conditional independence plays a central role in both statistical theory and applied fields such as causal inference. It forms important theoretical concepts such as sufficiency [**9**]. Meanwhile, the conditional independence relationship is a key factor in applications such as social network analysis [**13**]. Another application of the conditional Independence test is building directed acyclic graphs. Having valid conditional Independence testing methods is crucial in methods such as the PC algorithm [**39**]. Shah and Peters [**38**] have proved the discouraging fact that for continuous (X, Y, Z), no test will have any power if the test controls type-I error for all distributions where $X \perp\!\!\!\perp Y|Z$. This explains why despite huge efforts, there is no one testing method which can be applied to all circumstances. Existing methods [**22**] have primarily been focusing on the case when both X and Y are continuous. Sen et al. [**37**] provided an example of applying powerful machine learning classifiers to the challenging conditional independence problem. Hence, first, we would like to develop a testing method targeting the case in which X and Y are categorical while Z is continuous. Then, motivated by Sen et al. [**37**]'s idea, we would like our method to harness the power of the state of art machine learning tools.

With these goals, we combine neural networks with the Ising model / Boltzmann machine to represent a rich class of conditional distributions. We use the score test statistic and a score-based

bootstrap [**20**] algorithm to generate the p-value. This method provides a computationally efficient alternative to the generalized likelihood ratio test (GLRT) with the non-parametric bootstrap to generate p-values. We show that this method outperforms previously proposed computationally efficient methods, and is competitive to the computationally infeasible GLRT. We prove theoretical guarantees to back up these empirical observations.

The next two projects are applied data analysis on COVID-19 data. The coronavirus disease 2019 (COVID-19) pandemic has caused a worldwide impact on everyone's life. Healthy Davis Together (HDT) was an initiative to mitigate the impact of COVID-19 with a joint effort from the city of Davis and the University of California, Davis. I worked under HDT on the wastewater project (Chapter 3) and the college reopening analysis project (Chapter 4).

The wastewater project (a collaboration with the Wastewater team [**36**]) aims to provide a long-term wastewater-based solution for monitoring the COVID-19 outbreak. Wastewater monitoring serves as a useful tool in epidemiology to complement the clinical testing for managing COVID-19 [**36**]. Since wastewater monitoring does not require active participation from the public, it may reduce the sampling bias from the clinical data [**36**]. The wastewater team actively monitored the qPCR data of COVID-19 RNA using the qPCR technique. One challenge the wastewater team faced is how to handle the qPCR nondetects in the data. Both simple imputation (use a single value to impute all nondetects) and censoring (dropping nondetects) may also severe bias in the analysis [**36**]. To remedy the bias, we applied a Bayesian Ct value imputation method via the EM-MCMC algorithm wrapped in a user-friendly API. The algorithm is able to produce Ct values matching the overall trend of the clinical data and has a stronger correlation with the clinical data when compared with existing methods [**36**].

The final project at HDT is measuring the impact of college reopening on the COVID-19 outbreak level in their home county. Prior study [**25**] suggested that the opening of universities has led to super-spreader events, with a significant rise in confirmed cases reported by the universities. There are many factors that may influence the increase in confirmed cases, namely transmission rates, improved asymptomatic testing rates, or case importation due to the return of students to campus from alternative living situations during the summer. Is the return of students to campus to blame for the significant rise in COVID-19 cases? We examine the effect of the return of students

by matching the college counties to non-college counties that are within the same state and have a similar percentage of seniors (an important demographic variable for COVID-19 rates). Could the increase in cases actually be spurious, and a false alarm due to noisy case reporting? We cast this question as a multiple hypothesis testing problem and define a notion of hotspots that controls the false discovery rate. Third, are these hotspots associated with how the colleges reopened, such as in-person classes or the testing availability on campus? We investigate in greater detail a selection of colleges within the college counties and categorize them in terms of their COVID-19 mitigation measures. Then we test for associations between the measures and hotspot status.

CHAPTER 2

Conditional Independence Test with Neural Network

2.1. Introduction

Given a triplet of random vectors (X, Y, Z), we use $X \perp\!\!\!\perp Y|Z$ to denote X being conditionally independent of Y given Z. The central question of this paper is testing

$$H_0 : X \perp\!\!\!\perp Y|Z \text{ versus } H_1 : X \not\perp\!\!\!\perp Y|Z. \tag{2.1}$$

Conditional independence is a foundational concept with various applications in the both theoretical and applied field of statistics. From a theoretical perspective, conditional independence provides a unified language to describe concepts and phenomena such as sufficiency and Simpson's paradox [**9**]. On the other hand, conditional independence testing plays a cardinal role in causal inference as we will demonstrate. Demands of determining causal relationships have been soaring in fields like genetics and machine learning. Geneticists have been constructing gene regulatory networks from gene expression data to understand the causal relationship among genes [**50**]. In machine learning, researchers and practitioners are drawn to causality-driven machine learning models [**15**, **30**] which have more consistent predictive performance over time thanks to their robustness to spurious correlations. Directed acyclic graphs (DAG) are often fruitfully used to represent causal structures in both applications. Unless known from prior knowledge, DAGs need to be learnt from the data and this opens the gateway of structural learning. Some widely adopted methods in structural learning are the constraint-based approaches such as the PC algorithm [**39**] which relies on testing conditional independence to discover the skeleton of the graph. Hence, having valid and flexible conditional independence test methods adds powerful weapons to researchers' causal inference quivers.

In other applications, such as social network analysis [**13**], instead of causal relationships, correlation and condition independence themselves are of interest. Many existing methods tackle the

problem with an undirected graphical model (UGM). Like DAGs, researchers often have to learn the structure of the UGM before inferring conditional independence. One popular approach is specifying the full joint-likelihood with $l1$-regularization [**11**]. Gaussian model or Ising / Potts model are often used to model the joint distribution when all variables in the graph are continuous or categorical respectively. When the Gaussian model is assumed, we may use the zero entries of the inverse covariance matrix to characterize conditional independence among variables. However, as Neykov et al. [**27**] pointed out that departure from normality may lead to erroneous conclusions as when data is not multivariate normal, zero partial correlation is not equal to conditional Independence. In addition, both the Gaussian and the Ising / Potts models are restrictive on the type of variables they can model. For example, neither are suitable for count data or strictly positive continuous random variables. Yang et. al. [**48**] proposed a method to model the graph with distributions from univariate exponential family distributions. Still, with the advances in machine learning and computing power, we start to encounter unconventional data types such as text and images. Modeling distributions of them can be challenging as illustrated by copious language models available [**31**]. In addition, if we are only interested in a single or subset of conditional independence relationships which a graph can encode, we would like to test the specific conditional relationship directly rather than estimating the entire graphical model. This again calls for conditional independence test methods.

In summary, conditional Independence testing has applications in various fields. We will review the existing literature on non-graphical conditional Independence test methods in the next section. Readers who would like more exposure to graphical models and structural learning should check the excellent survey paper [**11**] on the subjects by Drton and Maathuis.

2.1.1. Related Work. We define the lowercase $p(.)$ as the probability density function (PDF) and the uppercase $P(.)$ as the probability mass function (PMF). So $p(X, Y|Z)$ is the density of the conditional distribution of $X, Y|Z$, and $P(X, Y|Z)$ is the conditional PMF.

Shah and Peters [**38**] prove that for continuous (X, Y, Z), no test will have any power if the test controls type 1 error for all distributions where $X \perp\!\!\!\perp Y|Z$. Neykov et al. [**27**] provide a more refined proof for the continuous case and prove the theorem for discrete X and Y. Their insight is that the culprit of the hardness of CI testing is the continuity of Z and the complexity of the space of all conditionally dependent distributions. At a high level, one can consider conditional independent

distributions of X and Y given Z are dense in conditional dependent distributions with respect to the Wasserstein distance. Because of the hardness of the CI problem, people have developed different assumptions and methods to make CI testing possible for certain scenarios.

Some of the methods we review here can be found in Li and Fan's survey paper [**22**] which provides a more detailed review with an empirical study. The first type of test is based on the discretization of Z. Huang [**17**] proposed a test statistic which is the expected value of maximal nonlinear conditional correlation between X and Y at selected Z values within the support. The method assumes X, Y, and Z are all continuous and generates the p-value based on the local bootstrap method based on the empirical CDF of Z, $X|Z$, and $Y|Z$. Neykov et al.'s [**27**] method is also based on the discretization of Z (and X and Y for continuous X and Y). They targeted cases when Z is continuous on $[0, 1]$ and X, Y are either discrete or continuous on $[0, 1]$. The test first discretizes Z, computes U-statistics on each bin, and generates the p-value with data permutation within each bin. In essence, methods based on discretization of Z only test $X \perp\!\!\!\perp Y|$ bin(Z). X being conditionally independent of implies $X \perp\!\!\!\perp Y|$ bin(Z), but not necessarily vice versa.

Kernel-based methods are primarily designed for continuous X, Y and have strong empirical performance in empirical studies [**22**]. The kernel conditional independence test (KCIT) by Zhang et al. [**49**] "essentially tests for zero Hilbert-Schmidt norm of partial cross-covariance operator" [**35**]. Despite computing its p-value through the asymptotic distribution of test statistics, it suffers from high computation cost because it needs to compute a matrix inverse "which scale strictly greater than $O(N^2)$" [**22**]. Zhang and Visweswaran [**40**] have developed a randomized conditional independence test and the randomized conditional correlation test to approximate KCIT to solve the scalability problem of KCIT.

Another category of tests targeting continuous X and Y are regression-based methods. Shah and Peter [**38**] proposed the generalised covariance measure which measures the correlation between residuals of regressing X on Z and Y on Z. The p-value is available through the asymptotic distribution of the test statistic. Metric-based testing methods exploit the fact that conditional distribution of $X, Y|Z$ factorizes under H_0. The conditional distance Independence test (CDIT) developed by Wang [**45**] measures the distance between $P_{X,Y|Z}$ and $P_{X|Z} * P_{Y|Z}$ through the conditional characteristic function. Despite providing theories of asymptotic of their test statistic,

authors still recommend using the nearest neighbor bootstrap method to generate the p-value due to computational difficulties. The nearest neighbor bootstrap method generates bootstrap samples by permuting X between two samples if their Z's are close. The test is applicable for both continuous and discrete X and Y. Runge [**35**] proposes a test statistic based on conditional mutual information

$$\int_Z \int_Y \int_X \log \frac{p(x, y|z)}{p(x|z)p(y|z)} dP_{X,Y,Z}.$$

The p-value can be generated by repeatedly calling the nearest neighbor bootstrap method and computing the test statistic on Bootstrap samples. Runge specifically mentions that the test is designed for continuous responses and the test exceeds the nominal level in simulations on binary X and Y in their simulations. The CCIT method proposed by Sen et al. [**37**] also uses the nearest-neighbour method to simulate data under H_0. However, instead of using a divergence metric, they train a classifier to distinguish if a sample is from the original data or the simulated data. If the classifier can perform better than random guessing, one can conclude the original violates H_0. Sen et al. only consider continuous X, Y in their paper, but the algorithm can be applied to discrete X and Y without adjustments.

Last but not least, there is the conditional randomized test (CRT) introduced by Candès et al. [**6**]. The CRT is more like a testing framework in the sense that it provides a general p-value generating mechanism allowing users to pick their own test statistic. The framework assumes that the conditional distribution of $X|Z$ is either known or can be learnt extremely well from prior data. The knowledge of the distribution of $X|Z$ allows one to generate bootstrap X^* and (X^*, Y, Z) has the same distribution as the original sample (X, Y, Z) under H_0. Then, the p-value can be produced with the bootstrap (X^*, Y, Z). Berrett et al. [**4**] propose the the conditional permutation test – a variant of the CRT. Instead of simulating X^* using the conditional distribution, they permute X from the original sample and use the conditional distribution of $X|Z$ to determine the permutation. Katsevich and Ramdas [**19**] create a most powerful test statistic for the CRT framework by exploiting the conditional validity of CRT and reducing the hypotheses to point null versus point alternative. Specifically, they assume the true $p^*(X|Z)$ is known and fix alternative

distributions $\bar{p}(Z), \bar{p}(Y|X,Z)$. Their original hypotheses are

$$H_0 : (X,Y,Z) \sim p(Z)p^*(X|Z)p(Y|Z) \text{ versus } H_1 : (X,Y,Z) \sim \bar{p}(Z)p^*(X|Z)\bar{p}(Y|X,Z) \tag{2.2}$$

for some $p(Z), p(Y|Z)$. H_0 is composite. By conditioning Y, Z, they change hypotheses to for each i,

$$H_0 : X_i|Y_i, Z_i \sim p^*(X|Z) \text{ versus } H_1 : X_i|Y_i, Z_i \sim p^*(X|Z)\frac{\bar{p}(Y_i|X_i, Z_i)}{\bar{p}(Y_i|Z_i)} \tag{2.3}$$

for given $\bar{p}(Y_i|Z_i)$'s. Since both H_0 and H_1 are points, the likelihood ratio test statistic $\frac{\bar{p}(X,Y|Z)}{\bar{p}(X|Z)}$ is most powerful by Neyman-Pearson Lemma. In practice, one has to learn $\bar{P}(X,Y|Z), \bar{P}(X|Z)$ on a training set and compute the likelihood ratio test statistic on a test set. We will refer to the likelihood ratio test statistic as the "Ising MP test statistic" in the remaining sections of the paper.

2.1.2. Summary of Results. As we review related works, we notice that most literature focuses on the case in which X and Y are continuous. The CDIT, CCIT, and CRT with the likelihood ratio test statistic are three tests we are aware of that can be applied to discrete X, Y without restrictions on the support of Z. Our method fills the gap by providing a conditional independence test when X and Y are discrete regardless of the continuity of Z. Our method uses the Ising model for binary X, Y and the restricted Boltzmann machine for multi-class X and Y. In both scenarios, we incorporate neural networks into the models so that they may model a rich class of distributions. Our method handles confounders (Z) of various forms such as text, images, and high-dimensional inputs. To reduce the computational cost, We adapt a score-based bootstrap method so that we only fit the model once. In our simulation study, we compare our method with the CCIT and CRT [1] method and show that our test statistic has superior type-I and II error control. We prove that the test statistic separates the null and alternative hypothesis asymptotically.

2.1.3. Organization. In Section 2.2, we first introduce the neural Ising model to model the distribution binary X and Y. Next, we discuss test statistics derived from the Ising model. Section 2.3 covers the Bootstrap procedure to generate the p-value. We review the wild bootstrap method

[1] We only include CRT when generating the RoC curve and omit it from further simulations because of its RoC curve.

and the score based bootstrap method [**20**]. Then we discuss how to apply bootstrap methods to our test statistics. In Section 2.4, we extend our method to multi-class X and Y through a restricted Boltzmann machine. Section 2.5 first illustrates simulation results of test statistics and bootstrap methods based on the Ising model. Then, it shows examples of applying the Ising model and restricted Boltzmann to real-world data. In Section 2.6, we provide theoretical results of the asymptotic separability of our test statistics.

2.2. Problem Settings and Methods

2.2.1. Neural Ising Model. Suppose that we observe X, Y, Z with $X, Y \in \{-1, 1\}$ and $Z \in \mathcal{Z}$. We allow the set $\mathcal{Z}$ to be flexible. It can be a set of continuous or discrete random variables in $\mathbb{R}^d$ or a set of images of text data. Let $J(Z) = (J_X(Z), J_Y(Z), J_{XY}(Z))$ be the output of a neural net. We further assume that the neural net has l layers and $J_\cdot$ share all layers except the lth layer. So

$$J(Z) = \left(W_X^\top h^{[l-1]}(Z), W_Y^\top h^{[l-1]}(Z), W_{XY}^\top h^{[l-1]}(Z)\right). \tag{2.4}$$

We define $h^{[0]}(Z) = Z$ and $h^{[l-1]}(Z)$ is a p-dimensional vector of the output of the $(l-1)$-st layer and $W_\cdot$'s are matrices of compatible dimensions. (2.4) implies that J_X, J_Y, and J_{XY} share weights in all layers except the final linear layers. We will model the joint distribution of $X, Y|Z$ with a log-likelihood of

$$\log P(X, Y|Z) = -J_X(Z) \cdot X - J_Y(Z) \cdot Y - J_{XY}(Z) \cdot XY - \psi(J(Z)). \tag{2.5}$$

where

$$\psi(J(Z)) = \log \sum_{(X,Y)} \exp\{-J_X(Z) \cdot X - J_Y(Z) \cdot Y - J_{XY}(Z) \cdot XY\}$$

The likelihood without the underlying neural network is often known as the Ising model. Define $\mathcal{P}$ be the set of all joint distributions of $X, Y|Z$ with the aforementioned likelihood and $\mathcal{P}_0 = \{P \mid P \in \mathcal{P}, J_{XY}(Z) = 0\}$. We refer to models in $\mathcal{P}$ as full models and models in $\mathcal{P}_0$ as reduced models. We test (2.1) under model (2.5):

$$H_0 : W_{XY} = 0 \text{ versus } \ H_1 : W_{XY} \neq 0.$$

Note this is a weaker hypothesis than hypotheses in (2.1) because $W_{XY} = 0$ implies conditional Independence but not vice versa. So our method has inflated type-I error for some distributions in H_0, but this is expected based on Shah and Peter's result.

We will use θ to denote W_{XY} from now on as it is the primary parameter of interest for our problem. We define θ_0 to be the parameter under H_0 and $\theta_0 = 0$ based on our null hypothesis.

2.2.2. Test Statistics and Model Fitting.

2.2.2.1. *Score Test Statistic.* We adopt the notation introduced in Kline and Santos' paper [**20**]. To compute the score test statistic, we first fit a reduced model $\hat{P}$ by solving the following optimization problem:

$$\hat{P}_0 = \arg\min_{P \in \tilde{\mathcal{P}}_0} L_n = \arg\min_{P \in \tilde{\mathcal{P}}_0} -\frac{1}{n}\sum_{i=1}^{n} \log P(X_i, Y_i | Z_i) = \arg\min_{P \in \tilde{\mathcal{P}}_0} \frac{1}{n}\sum_{i=1}^{n} l_i. \tag{2.6}$$

$\tilde{\mathcal{P}}_0$ is a subset of $\mathcal{P}_0$ and it contains models with $J(Z)$ follows a fixed set of architectures. We define $\tilde{P}$ similarly. Let $\hat{\theta}$ be the maximum likelihood estimator of θ_0,

$$D_i = (X_i, Y_i, h_i^{[l-1]}), \boldsymbol{\Sigma}(\theta) = E[s(D,\theta)s(D,\theta)^\top], \boldsymbol{\Sigma}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} s(D_i,\theta)s(D_i,\theta)^\top.$$

The test statistic has the quadratic form $G_n = T_n^\top T_n$ for a vector-valued T_n.

$$T_n = \left(\boldsymbol{A}_n(\hat{\theta})\boldsymbol{\Sigma}_n(\hat{\theta})\boldsymbol{A}_n(\hat{\theta})^\top\right)^{-\frac{1}{2}} S_n(\hat{\theta}),\ S_n(\hat{\theta}) = \boldsymbol{A}_n(\hat{\theta})\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(D_i,\hat{\theta}).$$

In this section, the p x 1 vector $s_i(\theta) = \nabla_\theta l_i$ which is the gradient of l_i with respect to θ. $\boldsymbol{\Sigma}(\theta)$is the covariance matrix of $\{s(D_i,\theta)\}_{i=1}^n$ and $\boldsymbol{\Sigma}_n(\theta)$ is the sample covariance matrix. $\boldsymbol{A}_n(\theta)$ is the inverse of the p x p matrix $\boldsymbol{H}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta^2 l_i$. $\boldsymbol{H}_n(\theta)$ is also known as the observed Fisher information. We denote the Fischer Information $E[\nabla_\theta^2 l]$ as $\boldsymbol{H}(\theta)$. A straightforward calculation will show that $\forall i \in \{1,\ldots,n\}$ and $j,k \in \{1,\ldots,p\}$,

$$\frac{\partial l_i}{\partial \theta_j} = h_j^{[l]}\left(\frac{2(e^{2(J_{X_i}+J_{Y_i})}+1)}{e^{2(J_{X_i}+J_{Y_i})}+e^{2(J_{X_i}+J_{XY_i})}+e^{2(J_{Y_i}+J_{XY_i})}+1} - X_iY_i - 1\right)$$

and

$$\frac{\partial^2 l_i}{\partial \theta_j \theta_k} = -\frac{8h_j^{[l]}h_k^{[l]}(\cosh(2J_{X_i})+\cosh(2J_{Y_i}))e^{2(J_{X_i}+J_{Y_i}+J_{XY_i})}}{\left(e^{2(J_{X_i}+J_{Y_i})}+e^{2(J_{X_i}+J_{XY_i})}+e^{2(J_{Y_i}+J_{XY_i})}+1\right)^2}$$

$\boldsymbol{A}_n(\hat{\theta})\boldsymbol{\Sigma}_n(\hat{\theta})\boldsymbol{A}_n(\hat{\theta})^\top$ is commonly known as the sandwich estimator of the variance of $s(h_{\cdot}^{[l-1]}, \theta)$. The appeal of the sandwich estimator is that it provides a consistent estimator of variance especially when the model is misspecified. However, when the sample size is not large, the sandwich estimator may be unstable and inefficient [**44**]. In addition, the sandwich estimator requires calculating a sample covariance matrix and its inverse. Thus, the computation can be expensive when the dimension of $s(h_{\cdot}^{[l-1]}, \theta)$ is high. The computation burden will become more evident when we discuss the bootstrap method. To that end, people often make the assumption that the model is well-specified, then $\boldsymbol{H}(\theta) = \boldsymbol{\Sigma}(\theta)$. Then, instead of the sandwich estimator, the observed Fisher information matrix $\boldsymbol{H}_n(\theta)$ becomes the new variance estimator and the test statistic takes the simpler form

$$T_n = \left(\boldsymbol{A}_n(\hat{\theta})\right)^{-\frac{1}{2}} S_n(\hat{\theta}),\ G_n = T_n^\top T_n. \tag{2.7}$$

Despite being less robust theoretically, (2.7) appears to have valid size control and high power against the H_1.

2.2.2.2. *KL Test Statistic.* We shall give a heads-up that the KL test statistic is studied only to serve as a baseline for the size and the power of a test statistic. We do not have a computational feasible way to compute the p-value with the KL test statistic. For two discrete probability distributions P and Q on A, we define

$$D_{KL}(P \parallel Q) = \sum_{a \in A} P(a) \log\left(\frac{P(a)}{Q(a)}\right)$$

which is the Kullback–Leibler (KL) divergence between P and Q.

We split the data into a training set and a test set, then fit model (2.5) on the training set and obtained the fitted parameters $\hat{J}(Z)$ by solving the following optimization problem

$$\hat{P} = \arg\max_{P \in \tilde{\mathcal{P}}} \frac{1}{n} \sum_{i=1}^{n} \log P(X_i, Y_i | Z_i).$$

Let $\hat{P}(X,Y|Z)$ be the likelihood with parameters $\hat{J}(Z)$ and $P_{\hat{P}}$ be the I-projection of $\hat{P}$ onto $\tilde{\mathcal{P}}_0$ with parameters $J(Z)$. The I-projection [**8**] $P^I(X,Y|Z) = \underset{Q\in\mathcal{P}}{\arg\min}\, D_{KL}(Q \parallel \hat{P})$. Under H_0,

$$\begin{aligned} D_{KL}(Q \parallel \hat{P}) &= E[X_i]\cdot(\hat{J}_X - J_X) + E[Y_i]\cdot(\hat{J_Y} - J_Y) + E[XY]\cdot \hat{J_{XY}} + \psi(\hat{J}(Z)) - \psi(J(Z)) \\ &= \tanh J_X \cdot (J_X - \hat{J}_X) + \tanh J_Y \cdot (J_Y - \hat{J}_Y) + \tanh J_X \cdot \tanh J_Y \cdot \hat{J}_{XY} + \psi(\hat{J}(Z)) - \psi(J(Z)) \end{aligned}$$

Simple algebra will show that under H_0, $E[X] = -\tanh J_X$. To find the $J(Z)$, we first set $J_{XY} = 0$, then solve the following first-order conditions

$$\begin{cases} \frac{\partial D_{KL}(Q\|\hat{P})}{\partial J_X} & = 0 \\ \frac{\partial D_{KL}(Q\|\hat{P})}{\partial J_Y} & = 0. \end{cases}$$

$$\Rightarrow \begin{cases} (\hat{J}_X - J_X)\cdot(\operatorname{sech} J_X)^2 + \tanh J_X + (\operatorname{sech} J_X)^2 \cdot \tanh J_Y \cdot \hat{J}_{XY} - \tanh(J_x) & = 0 \\ (\hat{J}_Y - J_Y)\cdot(\operatorname{sech} J_Y)^2 + \tanh J_Y + (\operatorname{sech} J_Y)^2 \cdot \tanh J_X \cdot \hat{J}_{XY} - \tanh(J_Y) & = 0 \end{cases}$$

$$\Rightarrow \begin{cases} (\hat{J}_X - J_X)\cdot(\operatorname{sech} J_X)^2 + (\operatorname{sech} J_X)^2 \cdot \tanh J_Y \cdot \hat{J}_{XY} & = 0 \\ (\hat{J}_Y - J_Y)\cdot(\operatorname{sech} J_Y)^2 + (\operatorname{sech} J_Y)^2 \cdot \tanh J_X \cdot \hat{J}_{XY} & = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \hat{J}_X - J_X + \tanh J_Y \cdot \hat{J}_{XY} & = 0 \\ \hat{J}_Y - J_Y + \tanh J_X \cdot \hat{J}_{XY} & = 0. \end{cases}$$

The set of equations doesn't have an analytic solution. Therefore, we solve it numerically with the fsolve function in the SciPy package [**43**]. The KL test statistic

$$T_{kl} = \frac{1}{n}\sum_{i=1}^{n} D_{KL}\left(P^I(X_i,Y_i|Z_i) \parallel \hat{P}(X_i,Y_i|Z_i)\right) \tag{2.8}$$

evaluated over a test set of size n.

2.2.2.3. *Hyperparameter Tuning.* Tuning hyperparameters is critical in adjusting the fit of a neural network model. Hyperparameters in our model include but are not limited to the number of hidden layers, the number of neurons in each hidden layer, the number of training epochs, and the learning rate. We first split the full data into the training set and validation set. Then we perform a randomized search on possible hyperparameter combinations by tracking the log-likelihood on

the validation set. To have a more intuitive way of measuring the fit, we recommend tracking the predictive performance as well. The prediction rule is

$$(X, Y) = \underset{(x,y)\in\{1,-1\}^2}{\arg\max} \hat{P}(X = x, Y = y|Z).$$

When classes are balanced, tracking the accuracy metric is sufficient. Otherwise, we also track precision and recall to evaluate the fit. Regardless of test statistics, we tune the model architecture using models in $\mathcal{P}$ because we do not know the ground truth and want our model to have the capacity to fit the data whether H_0 is true or not. Once the hyperparameter search is done, we fit the model with selected hyperparameters on the full data set. When computing the KL test statistic, we split the data into training, validation and test set. The training and validation sets are used for tuning and model fitting. The test statistic T_{kl} is computed on the test set.

2.3. Bootstrap P-Value

2.3.1. Score Based Wild Bootstrap. The score based bootstrap method [**20**] is based on the wild bootstrap method developed by Wu [**47**] and Liu [**23**]. So we shall begin the section with a high-level overview of the wild bootstrap method. Consider the linear model

$$Y_i = X_i^\top \beta_0 + \epsilon_i,\ i = 1, \ldots, n$$

with $Y_i,\ \beta_0 \in \mathbb{R}$ and $X_i \in \mathbb{R}^p$. Let $\hat{\beta}$ be the least square estimator of β_0 and $e_i = Y_i - X_i^\top \hat{\beta}$ be the residual. The wild bootstrap method produces B bootstrap samples

$$\{Y_i^b, X_i^b\}_{i=1,b=1}^{n,B} \qquad Y_i^b = X_i^\top \hat{\beta} + \epsilon_i^b \qquad \epsilon_i^b = W_i^b * e_i.$$

$\forall b \in \{1, \ldots, B\}$, $\{W_i^b\}_{i=1}^n$ are independent of $\{Y_i, X_i\}_{i=1}^n$ and has mean 0, variance 1. For example, the standard normal distribution is one candidate to generate $\{W_i^b\}_{i=1}^n$. On each bootstrap sample, the method asks for refitting the linear model and computing a bootstrap least square estimate $\hat{\beta}^b$. Let $\hat{\beta}^*$ denote the random variable of these bootstrap least square estimates. Condition on $\{Y_i, X_i\}_{i=1}^n$, the distribution of $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ is an empirical estimate of the distribution of $\sqrt{n}(\hat{\beta} - \beta_0)$. One may use the former to perform inference on β_0, e.g. constructing confidence intervals.

The vanilla wild bootstrap method requires the refitting of the model on each bootstrap sample. The refitting step may take a considerable amount of time even in linear regression, let alone a

more computationally challenging model. This leads us to the score based bootstrap (2.9) which produces a bootstrap distribution of estimators without refitting the model. One key observation Kline and Santos made is that "... the residuals only influence the limiting distribution of the OLS estimator through the score". Indeed, we have

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta_0) &= \sqrt{n}\left(\sum_{i=1}^n X_i X_i^\top\right)^{-1}\left(\sum_{i=1}^n X_i Y_i\right)-\beta_0 \\ &= \sqrt{n}\left(\sum_{i=1}^n X_i X_i^\top\right)^{-1}\left[\sum_{i=1}^n X_i(X_i^\top\beta_0+\epsilon_i)-\left(\sum_{i=1}^n X_i X_i^\top\right)\beta_0\right] \\ &= H_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i\epsilon_i\end{aligned}$$

where $H_n = \frac{1}{n}X_iX_i^\top$. Note that H_n is also the Hessian of the least square loss and $X_i\epsilon_i$ is the score of the loss evaluated at the true parameter β_0. Similarly, one can show that

$$\sqrt{n}(\hat{\beta}^*-\hat{\beta}) = H_n^{-1}\sum_{i=1}^n X_i\epsilon_i^* = H_n^{-1}\sum_{i=1}^n X_i e_i W_i^*.$$

Therefore, in order to simulate the empirical distribution of $\sqrt{n}(\hat{\beta}^*-\hat{\beta})$, one could simply perturb the score evaluated at $\hat{\beta}$ instead of refitting the linear model as prescribed in the original wild bootstrap method.

Kline and Santos extend the idea of perturbing scores to other models and develop a score based bootstrap method which can produce the p-value for a large class of tests. Specifically, they require the test statistic G_n to be the quadratic form $T_n^\top T_n$ and

$$T_n = \left(\boldsymbol{A}_n(\hat{\theta})\boldsymbol{\Sigma}_n(\hat{\theta})\boldsymbol{A}_n(\hat{\theta})^\top\right)^{-\frac{1}{2}} S_n(\hat{\theta})+o_p(1),\ S_n(\hat{\theta}) = \boldsymbol{A}_n(\hat{\theta})\frac{1}{\sqrt{n}}\sum_{i=1}^n s(D_i,\hat{\theta}).$$

D_i is a random vector containing both dependent and independent variables. $s(D_i,\hat{\theta})$ is the score vector of dimension p and $\boldsymbol{A}_n(\hat{\theta})$ is an r by p matrix. In the linear model example, $\boldsymbol{A}_n(\hat{\theta})$ is the inverse of the Hessian. To compute bootstrap test statistic G_n^*, one may replace T_n with T_n^*, where

$$T_n^* = (\boldsymbol{A}_n(\hat{\theta})\boldsymbol{\Sigma}_n^*(\hat{\theta})\boldsymbol{A}_n(\hat{\theta})^\top)^{-\frac{1}{2}} S_n^*(\hat{\theta})+o_p(1),\ S_n^*(\hat{\theta}) = \boldsymbol{A}_n(\hat{\theta})\frac{1}{\sqrt{n}}\sum_{i=1}^n s^*(D_i,\hat{\theta})W_i^*. \tag{2.9}$$

W_i is the noise with mean 0 and variance 1. $\boldsymbol{\Sigma}_n^*(\hat{\theta})$ is the sample covariance matrix of $\{s^*(D_i,\hat{\theta})\}_{i=1}^n$.

2.3.2. Generate the P-value with the Score Based Bootstrap. We adapt the score based wild bootstrap method to obtain the p-value for the score test statistic introduced in Section 2.2.2.1. Algorithm 1 provides a step-by-step description of the procedure. Recall in Section 2.2.2.1, we mentioned that the sandwich estimator has a higher computation burden. The point should be more evident after the introduction of Algorithm 1, as the sandwich variance estimator requires re-computing the sample covariance matrix of the scores and its inverse for each bootstrap test statistic.

Algorithm 1 Score Test Statistic P-Value

Input: $\{X_i, Y_i, Z_i\}_{i=1}^n$, model architecture, variance estimator type, number of bootstrap trials (B).

Fit a $\hat{P} \in \tilde{\mathcal{P}}_0$ based on model architecture.

Produce the output of the last hidden layer $\{h_i^{[l-1]}\}_{i=1}^n$ using $\hat{P}$.

Calculate the scores $\{s(D_i, \hat{\theta})\}_{i=1}^n$ based on $\hat{P}$, where $D_i = (X_i, Y_i, h_i^{[l-1]})$.

Compute $\boldsymbol{\Sigma}_n(\hat{\theta})$ which is the sample covariance matrix of $\{s(D_i, \hat{\theta})\}$.

Calculate $\boldsymbol{A}_n = H_n^{-1}$. H_n is the Hessian based on $\hat{P}$.

Compute $S_n(\hat{\theta}) = \boldsymbol{A}_n(\hat{\theta}) \frac{1}{\sqrt{n}} \sum_{i=1}^n s(D_i, \hat{\theta}))$.

if variance estimator type is the sandwich estimator **then**

$T_n = \left(\boldsymbol{A}_n(\hat{\theta}) \boldsymbol{\Sigma}_n(\hat{\theta}) \boldsymbol{A}_n(\hat{\theta})^\top\right)^{-\frac{1}{2}} S_n(\hat{\theta})$

else

$T_n = \left(\boldsymbol{A}_n(\hat{\theta})\right)^{-\frac{1}{2}} S_n(\hat{\theta})$

end if

Calculate the test statistic $G_n = T_n T_n^\top$.

for $b \in \{1 \ldots, B\}$ **do**

Generate i.i.d. W_i^b's with $E[W_i] = 0$, $Var(W_i) = 0$ and $i \in \{1, \ldots, n\}$.

Calculate perturbed scores $\{W_i^b s(D_i, \hat{\theta})\}_{i=1}^n$.

Produce perturbed scores $\{s^b(D_i, \hat{\theta}) = w_i * s_i(\theta_0)_{i=1}^n\}$.

Compute $S_n^b(\hat{\theta}) = \boldsymbol{A}_n(\hat{\theta}) \frac{1}{\sqrt{n}} \sum_{i=1}^n s^b(D_i, \hat{\theta}))$.

if variance estimator type is the sandwich estimator **then**

Compute $\Sigma_n^b(\hat{\theta})$ which is the sample covariance matrix of $\{s^b(D_i, \hat{\theta})\}_{i=1}^n$.

$T_n^b = \left(\boldsymbol{A}_n(\hat{\theta}) \Sigma_n^b(\hat{\theta}) \boldsymbol{A}_n(\hat{\theta})^\top\right)^{-\frac{1}{2}} S_n^b(\hat{\theta})$

else

$T_n^b = \left(\boldsymbol{A}_n(\hat{\theta})\right)^{-\frac{1}{2}} S_n^b(\hat{\theta})$

end if

$G_n^b = \left(T_n^b\right)^\top T_n^b$.

end for

Return: $\frac{1}{B} \sum_{i=1}^B \mathbb{1}\{G_n^b > G_n\}$

Because we have not discovered a way to decompose the KL test statistic into the form described in Section 2.3.1, we cannot apply the score based bootstrap method to the KL test statistic. Therefore, we do not have a method to compute the bootstrap p-value within a reasonable time. However, in principle, we could produce a p-value KL test statistic by generating bootstrap samples under H_0 with a $\tilde{P} \in \tilde{\mathcal{P}}_0$ learnt on the data. Algorithm 2 describes how one would implement the idea. Readers who are not interested in the thought process may safely skip to the next section.

Algorithm 2 KL Test Statistic P-Value

Input: $\{X_i, Y_i, Z_i\}_{i=1}^n$, model architecture, $\tilde{P}$_method, number of bootstrap trials (B).

Split the data into a training set and a test set.

Fit a $\hat{P} \in \tilde{\mathcal{P}}$ on the training set based on model architecture.

Compute the KL test statistic T_{kl} described in Section 2.2.2.2 on the test set.

Produce a $\tilde{P} \in \tilde{\mathcal{P}}_0$ based on $\tilde{P}$_method on the training set.

for $b \in \{1 \ldots, B\}$ **do**

Generate $\{X_i^b, Y_i^b\}_{i=1}^n$ using $\tilde{P}$ condition on $\{Z_i\}_{i=1}^n$ in the training set.

Fit a $\hat{P}^b \in \mathcal{P}$ based on model architecture and $\{X_i^b, Y_i^b, Z_i^b\}_{i=1}$.

Compute the KL test statistic T_{kl}^b based on $\hat{P}^b$ on the test set.

end for

Return: $\frac{1}{B}\sum_{i=1}^B \mathbb{1}\{T_{kl}^b > T_{kl}\}$

There are a few choices of how to learn $\tilde{P}$ appearing in Algorithm 2. The obvious one is fitting a model in $\tilde{\mathcal{P}}_0$ in addition to $\hat{P}$. To avoid refitting, one could either let $\tilde{P}$ be the I-projection of $\hat{P}$ onto $\tilde{\mathcal{P}}_0$, or let $\tilde{P}$ share parameter values with $\hat{P}$ but set $J_{XY} = 0$. Algorithm 2 is similar to Candes et al. [**6**] and Berrett et al. [**4**]'s methods. The difference is that their methods only estimate $P(X|Z)$ and use permutation to simulate data under H_0; whereas we estimate both $P(X|Z)$ and $P(Y|Z)$.

Ultimately, we didn't use Algorithm 2 to generate P-values, because refitting the model in each bootstrap iteration is too computationally expensive.

2.4. Restricted Boltzmann Machine for Categorical Responses

In this section, we discuss the extension of our method to multi-class X and Y. We focus on the case in which X and Y are categorical but not ordinal. Define e_i to be a one-hot vector of which the ith entry is 1 and 0 otherwise. E.g, if $e_2 \in \mathbb{R}^3$, $e_i = [0, 1, 0]^\top$. Suppose X has k_x classes and Y has k_y classes, then $X \in \{e_i : e_i \in \mathbb{R}^{k_x}\}$ and $Y \in \{e_i : e_i \in \mathbb{R}^{k_y}\}$. Define $\mathbf{1}^k \in \mathbb{R}^k$ to be a vector of all ones. We only consider the conditional distribution of the form

$$\log P(X, Y|Z) = X^\top J(Z)Y - \log \sum_{X,Y} \exp\{X^\top J(Z)Y\} \tag{2.10}$$

and $J(Z)$ is a k_x by k_y parameter matrix with each entries being a function of Z. We assume that the function is an l-layer neural network in which all outputs shared layers except the last layer. So $\forall i \in \{1, \ldots, k_x\}, j \in \{1, \ldots, k_y\}$,

$$J_{ij}(Z) = \Phi_{ij}^\top h^{[l-1]}(Z) \tag{2.11}$$

where $h^{[l-1]}(Z) \in \mathbb{R}^p$ is the output of the $l-1$th layer and $\Phi_{ij} = (\Phi_{ij1}, \ldots, \Phi_{ijp}) \in \mathbb{R}^p$ is the weights. We will explain why the restriction on network architecture exists during hypotheses formulation. Similarly, we only consider the conditional independent distribution of the form

$$\log P(X, Y|Z) = X^\top \gamma^x(Z) + \gamma^y(Z)^\top Y - \log \sum_X \exp\{X^\top \gamma^x(Z)\} - \log \sum_Y \exp\{\gamma^y(Z)^\top Y\}. \tag{2.12}$$

We impose restrictions like (2.11)– namely, $\forall i \in \{1, \ldots, k_x\}, j \in \{1, \ldots, k_y\}$,

$$\gamma_i^x(Z) = \alpha_i^\top h^{[l-1]}(Z),\ \gamma_j^y(Z) = \beta_j^\top h^{[l-1]}(Z), \alpha_i, \beta_j \in \mathbb{R}^p. \tag{2.13}$$

Define $\mathcal{P}^b$ be the set of all conditional distributions satisfying (2.10) and (2.11) and $\mathcal{P}_0^b$ be the set of all conditional distributions satisfying (2.12) and (2.13). Note under model (2.12), conditional

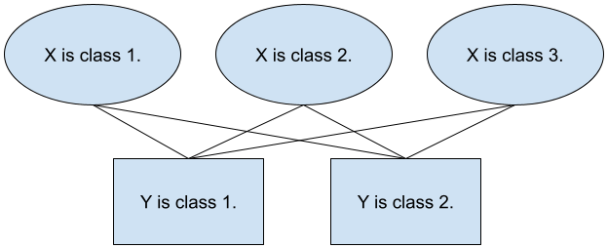


FIGURE 2.1. This is a graphical illustration of model (2.10) when ignoring the dependency on Z. In this example, X has three classes and Y has two classes. Each node is associated with a class and only one node in each row can equal 1 at a time. In $\mathcal{P}^b \setminus \mathcal{P}_0^b$, PMF is determined by edges (J_{ij}). Whereas, in $\mathcal{P}_0^b$, all edges are eliminated and the PMF is determined by nodes.

distributions of $X|Z$ and $Y|Z$ are two softmax models and

$$\begin{aligned} X^\top \gamma^x(Z) + \gamma^y(Z)^\top Y &= X^\top \gamma^x(Z)(\mathbf{1}^{k_y})^\top Y + X^\top \mathbf{1}^{k_x} \gamma^y(Z)^\top Y \\ &= X^\top [\gamma^x(Z)(\mathbf{1}^{k_y})^\top + \mathbf{1}^{k_x} \gamma^y(Z)^\top] Y \\ &= X^\top J(Z) Y. \end{aligned} \tag{2.14}$$

So model (2.12) is just a special case of model (2.10) with

$$J(Z) = \gamma^x(Z)(\mathbf{1}^{k_y})^\top + \mathbf{1}^{k_x} \gamma^y(Z)^\top \tag{2.15}$$

for some $\gamma^x(Z) \in \mathbb{R}^{k_x}$ and $\gamma^y(Z) \in \mathbb{R}^{k_y}$. Without the underlying neural network for $J(Z)$, the model specified in Equation (2.10) is commonly known as a restricted Boltzmann machine (RBM). Figure 2.1 illustrates RBM as an undirected graphical model.

2.4.1. Hypotheses Formulation. In this section, we will characterize conditional independence under the RBM model, then formulate hypotheses with the characterization. We begin with a proposition.

PROPOSITION 1. *Given $\mathcal{P}^b$ and $\mathcal{P}_0^b$, if $P(X, Y|Z) \in \mathcal{P}^b$ and $X \perp\!\!\!\perp Y|Z$ if and only if $P(X, Y|Z) \in \mathcal{P}_0^b$.*

PROOF. The backward direction is trivial. We will prove the forward direction. Since $P(X, Y|Z) \in \mathcal{P}^b$, without the normalizing constant, $\log P(X, Y|Z) = X^\top J(Z)Y$. If $X \perp\!\!\!\perp Y|Z$, $X^\top J(Z)Y$ factorizes. Let $J(Z)^x$ and $J(Z)^y$ be two k_x by k_y matrices. If $X^\top J(Z)Y$ factorizes,

$$X^\top J(Z)Y = \underbrace{X^\top J(Z)^x Y}_{a} + \underbrace{X^\top J(Z)^y Y}_{b}$$

and term (a) and (b) each only depends on one random variable. Without losing generality, let's assume term (a) doesn't depend on Y. If term (a) depends only on X, $J(Z)^x Y$ doesn't depend on Y. If $J(Z)^x Y$ doesn't depend on Y, $J(Z)^x Y = \gamma^x(Z)$ for some $\gamma^x(Z)$ doesn't depend on Y. Similarly, if term (b) depends only on Y, $X^\top J(Z)^y = \gamma^y(Z)$ and $\gamma^y(Z)$ is a vector that doesn't depend on X. If $J(Z)^x Y = \gamma^x(Z)$ and $X^\top J(Z)^y = \gamma^y(Z)$, $\log P(X, Y|Z)$ has the form in (2.12). So $\log P(X, Y|Z) \in \mathcal{P}_0^b$. □

Proposition 1 allows us to test conditional independence with the following hypotheses:

$$H_0 : P(X, Y|Z) \in \mathcal{P}_0^b \text{ versus } \quad H_1 : P(X, Y|Z) \in \mathcal{P}^b \setminus \mathcal{P}_0^b.$$

Testing $P(X, Y|Z) \in \mathcal{P}_0^b$ is equivalent to testing if

$$J(Z) = \gamma^x(Z)(\mathbf{1}^{k_y})^\top + \mathbf{1}^{k_x}\gamma^y(Z)^\top.$$

So H_0 is composite making applying conventional testing methods such as the score test challenging. We would like our hypothesis to be of the form $g(\Phi) = 0$ instead of $g(\Phi, Z) = 0$ for some parameter Φ.

We solve the second problem but at the cost of the power of the test. First, we need a few new definitions. Under (2.10), (2.11), we have

$$J(Z) = \begin{bmatrix} \Phi_{11}^\top h^{[l-1]}(Z) & \dots & \Phi_{1k_y} h^{[l-1]}(Z) \\ \vdots & & \\ \Phi_{k_x 1}^\top h^{[l-1]}(Z) & \dots & \Phi_{k_x k_y}^\top h^{[l-1]}(Z) \end{bmatrix}.$$

Let $\Phi = (\Phi)_{ijk} i \in \{1, \ldots, k_x\}$, $j \in \{1, \ldots, k_y\}, k \in \{1, \ldots, p\}$ be the parameter tensor. $\forall \Phi_{ijk} \in \mathbb{R}$, we define

$$\bar{r}_{i.k} = \frac{1}{k_y} \sum_{j=1}^{k_y} \Phi_{ijk}, \quad \bar{c}_{.jk} = \frac{1}{k_x} \sum_{i=1}^{k_x} \Phi_{ijk}, \quad \bar{m}_{..k} = \frac{1}{k_x k_y} \sum_{i=1}^{k_x} \sum_{j=1}^{k_y} \Phi_{ijk}$$

Define the linear operator, $P : \mathbb{R}^{k_x \times k_y} \to \mathbb{R}^{k_x \times k_y}. \forall i \in \{1, \ldots, k_x\}, j \in \{1, \ldots, k_y\}, k \in \{1, \ldots, p\}$

$$(P\Phi)_{ijk} = \Phi_{ijk} - \bar{r}_{i.k} - \bar{c}_{.jk} + \bar{m}_{..k}$$

Now we are ready for the following result

THEOREM 1. *$P(X, Y|Z) \in \mathcal{P}_0^b$ implies that $P\Phi = 0$.*

PROOF. Since $P(X, Y|Z) \in \mathcal{P}_0$,

(by (2.14))

$$J(Z) = \begin{bmatrix} \Phi_{11}^\top h^{[l-1]}(Z) & \ldots & \Phi_{1k_y} h^{[l-1]}(Z) \\ \vdots & & \\ \Phi_{k_x 1}^\top h^{[l-1]}(Z) & \ldots & \Phi_{k_x k_y}^\top h^{[l-1]}(Z) \end{bmatrix}$$

(by (2.15))

$$= \begin{bmatrix} (\alpha_1 + \beta_1)^\top h^{[l-1]}(Z)^\top & \ldots & (\alpha_1 + \beta_{k_y})^\top h^{[l-1]}(Z) \\ \vdots & & \\ (\alpha_{k_x} + \beta_1)^\top h^{[l-1]}(Z) & \ldots & (\alpha_{k_x} + \beta_{k_y})^\top h^{[l-1]}(Z) \end{bmatrix}$$

$$= \begin{bmatrix} (\alpha_{11} + \beta_{11})^\top h_1^{[l-1]}(Z) & \ldots & (\alpha_{1p} + \beta_{1p})^\top h_p^{[l-1]}(Z) & \ldots & (\alpha_{1p} + \beta_{k_y p})^\top h_p^{[l-1]}(Z) \\ \vdots & & & & \\ (\alpha_{k_x 1} + \beta_{11})^\top h_1^{[l-1]}(Z) & \ldots & (\alpha_{k_x p} + \beta_{1p})^\top h_p^{[l-1]}(Z) & \ldots & (\alpha_{k_x p} + \beta_{k_y p})^\top h_p^{[l-1]}(Z) \end{bmatrix}$$

Then $\forall i \in \{1, \ldots, k_x\}, j \in \{1, \ldots, k_y\}, k \in \{1, \ldots, p\}, \Phi_{ijk} \in \mathbb{R}$,

$$\begin{aligned}(P\Phi)_{ijk} &= \alpha_{ik} + \beta_{jk} - \frac{1}{k_y}\sum_{j=1}^{k_y}(\alpha_{ik} + \beta_{jk}) - \frac{1}{k_x}\sum_{i=1}^{k_x}(\alpha_{ik} + \beta_{jk}) + \frac{1}{k_x k_y}\sum_{i=1}^{k_x}\sum_{j=1}^{k_y}(\alpha_{ik} + \beta_{jk}) \\ &= -\frac{1}{k_y}\sum_{j=1}^{k_y}\beta_{jk} - \frac{1}{k_x}\sum_{i=1}^{k_x}\alpha_{ik} + \frac{1}{k_x k_y}\sum_{i=1}^{k_x}\sum_{j=1}^{k_y}(\alpha_{ik} + \beta_{jk}) \\ &= 0.\end{aligned}$$

□

We formulate our hypotheses as

$$H_0 : P\Phi = 0 \text{ versus } \; H_1 : P\Phi \neq 0.$$

Since Theorem 1 only has one direction, it is possible that distribution in $P^b \setminus P_0^b$ but $P\Phi = 0$. Therefore, our test may not have power against some distributions in H_1. Nevertheless, now our hypotheses are stated in terms of population parameters only and we can apply the score test and other parametric tests.

2.4.2. Score Test Statistic. To compute the score test statistic, we fit a Boltzmann machine of $X, Y|Z$ with

$$\Phi_{ijk} = \alpha_{ik} + \beta_{jk}, \quad \forall i, j, k.$$

We then flatten Φ into a vector and fill in the vector by rows. We denote the flattened parameter vector as θ. The test statistic still has the quadratic form $G_n = T_n^\top T_n$ for a vector valued T_n.

$$T_n = \left(\boldsymbol{A}_n(\hat{\theta})\boldsymbol{\Sigma}_n(\hat{\theta})\boldsymbol{A}_n(\hat{\theta})^\top\right)^{-\frac{1}{2}} S_n(\hat{\theta}), \; S_n(\hat{\theta}) = \boldsymbol{A}_n(\hat{\theta})\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(D_i, \hat{\theta}).$$

However, since we have the additional P in our hypotheses, we need to modify the test statistic to achieve a better convergence rate [2]. Specifically, $\boldsymbol{A}_n(\theta) = \dot{P}\boldsymbol{H}_n(\theta)^{-1}$ where $\dot{P}$ is the Jacobian matrix of $P\Phi$ with respect to Φ. Since P is a linear transformation, $\dot{P} = P$. For the variance estimator, we can still choose between the sandwich estimator and the Fisher information matrix

[2]For a formal justification, see Chapter 12.6.2 in Wooldridge [**46**].

by deciding whether to assume $H_n(\theta) = \mathbf{\Sigma}_n(\hat{\theta})$. Combine these changes and we have

$$T_n = \left(\dot{P}\boldsymbol{H}_n(\hat{\theta})^{-1}\dot{P}^\top\right)^{-\frac{1}{2}} S_n(\hat{\theta}),\ S_n(\hat{\theta}) = \dot{P}\boldsymbol{H}_n(\hat{\theta})^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(D_i, \hat{\theta}).$$

Algorithm 1 can still be applied to obtain the p-value.

2.5. Empirical Results

2.5.1. Simulations. We conducted simulation studies sensitivity and specificity of various test statistics and the size of associated tests. We generated $Z_i \in \mathbb{R}^3$ from the multivariate standard normal. Then we considered two conditional distributions of $X, Y|Z$.

(1) Ising Data
- $X, Y|Z \sim P$, where $P \in \mathcal{P}_0$ under H_0 and $P \in \mathcal{P}$ under H_1.
- We let $J(Z)$ be a 2-layer neural network with 100 hidden nodes.

(2) Mixture Data
- Under H_0, $P(X = 1|Z) = P(Y = 1|Z) = \frac{1}{1+\exp\{-||Z||_2\}}$.
- Under H_1, if $||Z||_2 < 1.53$, $X = -Y$. Otherwise, $X = Y$.

For each data-generating mechanism, we simulated data under 4 sample sizes (100, 500, 1000, 2000). Under each sample size and hypothesis (H_0 or H_1) combination, we generated 1000 data sets. In essence, the data generation setting can be represented by a triplet (e.g (Ising data, H_0, 100)). Note that under mixture data H_1, X but Y are conditionally dependent and marginally independent.

There are four test statistics based on the Ising model. They are Ising KL, Ising MP, Ising Score Sandwich, and Ising Score Fisher. Ising KL refers to test statistic (2.8). Ising Score Sandwich and Ising Score Fisher are test statistics introduced in Section 2.2.2.1 with corresponding variance estimators. Ising MP test statistic has the form $\Pi_{i=1}^{n}\frac{\hat{P}(Y_i|X_i,Z_i)}{\hat{P}(Y_i|Z_i)}$ which is introduced in Katsevich and Ramdas's paper [**19**]. When computing Ising KL and Ising MP test statistics, we reserve 10% of the data as the test data. We fit a model $\hat{P} \in \mathcal{P}$ on the training set and compute the test statistic on the test set. Other Ising test statistics are based on a fitted model $\hat{P}_0 \in \mathcal{P}_0$ and don't require a train-test split. On the Ising data, both $\hat{P}$ and $\hat{P}_0$ have the same architecture as the data

generating distribution P. So there is no model misspecification. On the mixture data, both $\hat{P}$ and $\hat{P}_0$ have a network with 2 hidden layers and 40 nodes in each layer.

We also include test statistics from three other tests for comparison. The Chi-square goodness of fit test (Naive Chi Sq) forms a contingency table of X and Y. The test statistic is

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}.$$

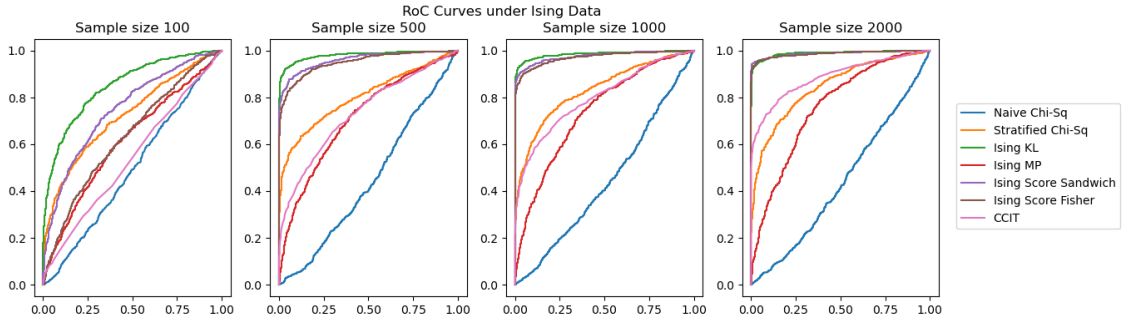
where O_{ij} is the observed cell count in the ith row and jth column of the table and E_{ij} is is the expected cell count in the ith row and jth column of the table. Although, the Chi-square goodness of fit tests for the independence of X and Y rather than the conditional independence, we include it to demonstrate the consequence of ignoring confounding factors Z. The stratified Chi-square test is an extension of the Chi-square goodness of fit tests in the sense that it performs a goodness of fit test on each stratum. In this study, we first run a k-means clustering algorithm with $k = 2$, then compute the test statistic

$$\sum_{c=1}^{2}\sum_{i=1}^{2}\sum_{k=1}^{2}\frac{(O_{cij}-E_{cij})^2}{E_{cij}}$$

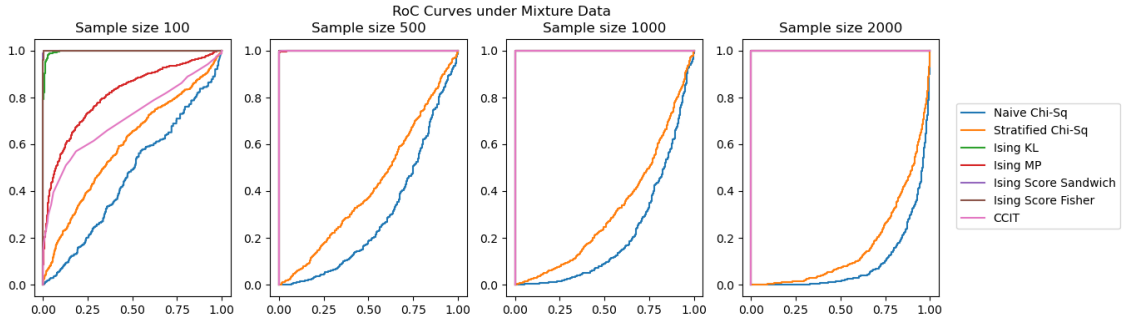
where O_{cij} is the observed cell count in the ith row and jth column of the table of cluster c. E_{cij} is defined accordingly.

The final test is the CCIT method in Sen et al.'s paper [**37**]. Although in the original paper, authors assume that both X and Y are continuous random variables, the method in principle can be applied to discrete X and Y. We use the code[3] provided by the author with default settings.

[3]Link: https://github.com/rajatsen91/CCIT



(A)

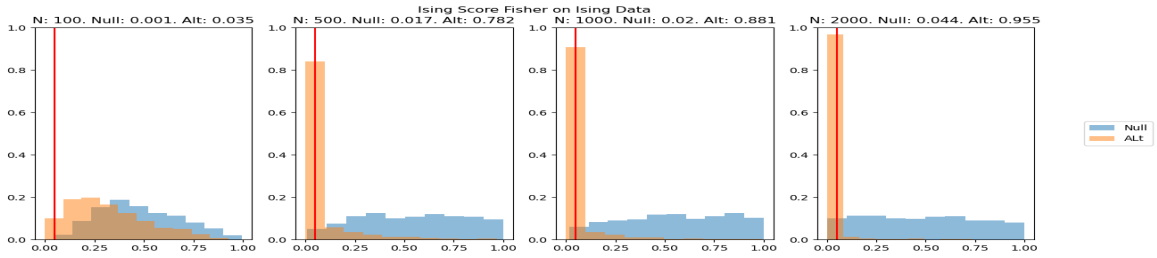


(B)

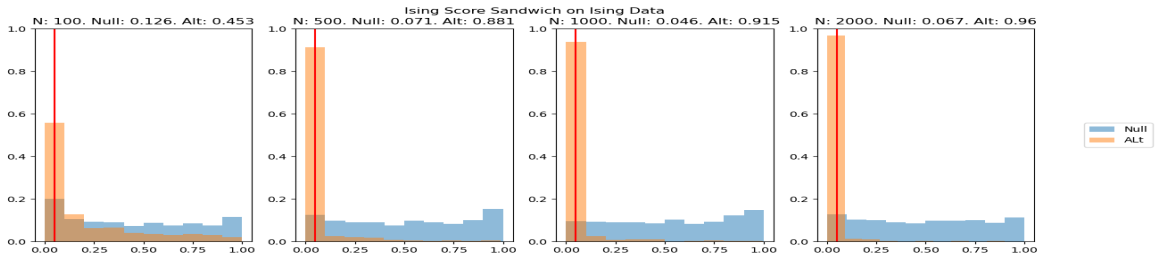
FIGURE 2.2. RoC curves of different methods fitted under the Ising Data (Top) and mixture data (bottom).

Figure 2.2 contains RoC curves of all test statistics under different settings. Under the Ising data, unsurprisingly, all test statistics based on the exact Ising model perform well, though Ising MP has considerably less power. Ising KL test statistic demonstrates the best capability to distinguish H_0 from H_1. As sample size increases, Ising Score Fisher and Ising Score Sandwich perform similarly compared to Ising KL. The Chi-square goodness of fit test can't separate H_0 from H_1, but it also doesn't pick up erroneous signals. The stratified version performs commendably well especially when the sample size is low. The k-means clustering probably produced reasonable strata. On the mixture data, the Chi-square goodness of fit test picks up the incorrect signal because of the construction of (Mixture data, H_1, $*$). Our simple stratification fails to account for confounders. On the other hand, our proposed test statistics have excellent powers and advantages over competing methods when the sample size is small.

Figure 2.3 and 2.4 shows the distribution of P-values produced by Algorithm 1. P-values of the Ising Score Fisher test seem to be conservative when the sample size is small. As the sample size grows, p-values under H_0 distribute more uniformly regardless of the data generating settings. In contrast, Ising Score Sandwich rejects H_0 more aggressively under both H_0 and H_1. The chance of rejecting H_0 is higher than the nominal level of the test under H_0 on the mixture data with given sample sizes. Since the Ising Score Fisher test is better at controlling the type 1 error and computationally less taxing (one less matrix to invert), we recommend it over the sandwich version.

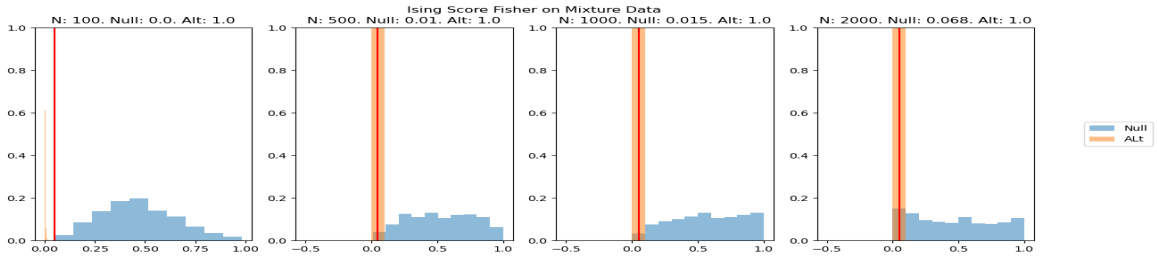


(A)

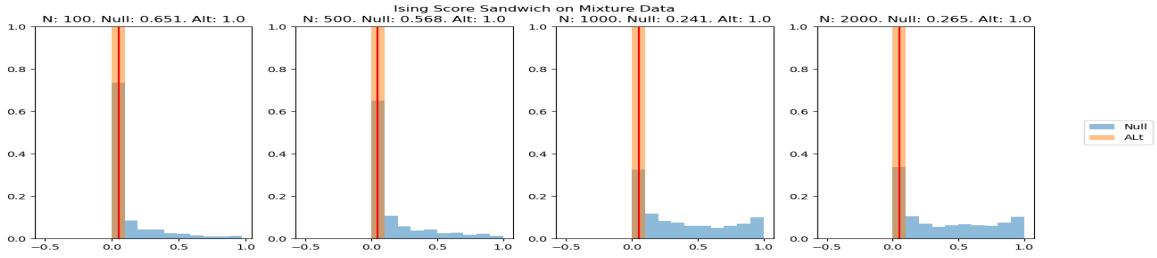


(B)

FIGURE 2.3. Histograms of P-values of Algorithm 1 on the Ising Data. The red line sits at 0.05. Numbers in each subtitle indicate the sample size and the rejection rate of H_0 at 0.05 level.



(A)



(B)

FIGURE 2.4. Histograms P-values of Algorithm 1 on the Mixture Data. The red line sits at 0.05. Numbers in each subtitle indicate the sample size and the rejection rate of H_0 at 0.05 level.

2.5.2. The Adult Data Set. For assessing the fairness of policies we typically identify protected attributes and determine if the effect of interest is dependent on the protected attributes. However, there may be "benign factors" such that if they explain this dependence then we can still say that the policy is fair. For example, when assessing the gender wage gap, we typically ask the question: do women make the same amount as men *for the same job*? The occupation, possibly in addition to resume, location, etc. are the benign factors in this case.

We apply Algorithm 1 with the variance estimator being the Fisher information to the Adult Data Set [**12**] to assess fairness. We study the question of whether income is independent of gender conditioning on factors such as the sector of jobs ("workclass"), degree level ("education"), occupation, native country, age, weight of each sample ("fnlwgt"), years of education ("education-num"), "capital-gain", "capital-loss" and hours-per-week. Since income and gender are binary random variables in the data set, we fit the Ising model which is designed for binary responses. The data set has already been split into a training set and a test set on the UCI Machine Learning

Repository. For simplicity, we only use the training set to implement Algorithm 1. To show that our model fits the data reasonably well, we report the predictive performance on the test set in Table 2.1. Since complicated architectures improve neither the likelihood nor the predictive performance, we let $J(Z)$ be simple linear functions of confounders Z. Namely, $J(Z) = (Z^\top W_X, Z^\top W_Y, Z^\top W_{XY})$. After selecting the architecture and other hyperparameters such as learning rate, we run Algorithm 1 and get a p-value of 0. So our method concludes that there is a gender pay gap after controlling for confounders such as age and educational background etc.

	Accuracy	Precision	Recall	F1
Sex (Female)	0.746	0.673	0.433	0.526
Income ($>$ 50k)	0.808	0.647	0.481	0.551

TABLE 2.1. Predictive Performance of $\hat{P} \in \mathcal{P}$. Metrics are based on minority classes ("Female" and "$>$ 50k").

2.5.3. The Cells Out of Sample Dataset. The adult data set provides a showcase for the binary Ising model with a simple linear relationship between Z and $J(Z)$. Now we illustrate an application of the Boltzmann machine model with neural networks through the Cells Out of Sample Dataset (COOS) [**24**]. COOS contains 132,209 images of mouse cells divided almost evenly into 7 classes. The data set has one training set and four test sets with various degrees of covariate shift. Covariate shift [**26**] is defined as

$$P(Y|X, \text{ training}) = P(Y|X, \text{ test}), P(X|\text{ training}) \neq P(X|\text{ test}).$$

In words, covariate shift means that the distribution of input data X shifts between the training set and the test set. The test set may contain X that is not seen in the training set (out-of-sample), therefore, the models' predictive performance on the test set may suffer [**24**] because of extrapolation when predicting labels on the test set.

Lu et al. [**24**] create images with covariate shift by taking images on different days and plates, etc. These changes should not alter the definitions of the cell class/label. So the relationship

between images and their labels shouldn't change because of the alternation in photo-taking techniques. However, in general, other types of distribution shifts may occur alongside the change of distribution of X. The covariate shift assumption

$$P(Y|X,\ \text{training}) = P(Y|X,\ \text{test})$$

does not necessarily hold. We would like to verify that there are no other types of distribution shift that happens with the covariate shift in the COOS data. The ability to answer this type of question may help researchers and practitioners to diagnose model performance drop going from training to production in the real world.

In essence, we test whether the covariate shift assumption holds. The covariate shift assumption can be tested with a conditional independence test. For the COOS data set, set

$$Z = \begin{cases} 0, & \text{if the sample is from the training set} \\ 1, & \text{if the sample is from the test set.} \end{cases}$$

Y to the image label, and X to the image. Then the hypotheses become

$$H_0 : P(Y|X, Z) = P(Y|X) \text{ versus } \ H_1 : P(Y|X, Z) \neq P(Y|X).$$

We create two experiments to test if our model can determine the existence of distribution shifts by mixing the training data with different test sets. In the first experiment, we use test set 1 which is a random holdout from the original training set. So test set 1 doesn't have covariate shift and as illustrated in Lu et al.'s paper [**24**], models show little signs of predictive performance loss. The best model only shows 0.4% increase in classification error compared to 1% on the training set. In the second experiment, we use test set 3 of which "images from 2 independent plates for each class, were reproduced on different days than the training dataset". Many models suffer the largest increase in classification error on the test set 3 with the smallest increase being 5.2%. Due to memory constraints, we limit the sample size to 5000 and only use images with labels 0, 1, and 2. Data are summarized in Table 2.2.

	Training Sample Prop	Image Class 0	Image Class 1	Image Class 2
Experiment 1	0.81	0.37	0.18	0.45
Experiment 2	0.56	0.4	0.31	0.29

TABLE 2.2. Summary of Labels in COOS Experiments. Training Sample Prop means the proportion of samples from the training set.

We fit an RBM model with $J(Z)$ being the neural network of the following architecture:

(1) A pre-trained VGG16 network without the top 3 layers from Tensorflow.

(2) Three fully connected layers with Exponential Linear Unit (ELU) function as activation functions.

(3) A softmax layer.

Each fully connected layer has the output dimension half of the input dimension. The dimension of the output of the final layer is 128. Again, to demonstrate the fit of our model, we report predictive performance on the validation set along with p-values.

	Data Set Label Accuracy	Image Label Accuracy	P-value
Experiment 1	0.767	0.973	0.364
Experiment 2	0.824	0.985	0.111

TABLE 2.3. Summary of Test Results.

In both experiments, our model has comparable performance against models in Lu et al.'s paper [**24**]. More importantly, our models fail to reject H_0 in both cases with a common level 0.05. We also include the CCIT method for comparison. The CCIT method uses the $J(Z)$ without the softmax layer as input and the predict the (X, Y) pair is from the original sample and a bootstrap sample. The prediction accuracy is 0.52 and 0.45 for test set 1 and 3 respectively. So, if we use 50% as the threshold, CCIT concludes that test set 1 does not satisfy the covariate shift assumption, while test set 3 satisfies the assumption.

2.6. Theories in the Linear Case

This section discusses theories of asymptotic separability of our test statistics under the Ising model. We assume all weights except W_{XY} are known. In other words, we assume that W_X, W_Y, and all weights in the hidden layer are assumed to be known. Hence, our theories are for the case when $J_{XY}(Z)$ is a linear function. Asymptotic separability means that the probability of type 1 and type 2 errors go to 0 asymptotically. We first introduce a few new notations and definitions most of which follow the convention in [**29**]. Recall we used θ to denote $W_{XY}^{[l]}$. We assume θ and $h^{[l-1]} \in \mathbb{R}^p$. In equation (2.6), we defined

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log P_\theta(X_i, Y_i|Z_i) = \frac{1}{n}\sum_{i=1}^{n} l_i(\theta).$$

and $L(\theta) = \log P_\theta(X, Y|Z)$. Define

$$\theta_0 = \underset{\{\theta: P(X,Y|Z)\in\mathcal{P}_0\}}{\arg\min} L(\theta).$$

So $\theta_0 = 0$, but we kept the notation θ_0 to remind the reader what it is conceptually. We will make the dependence between l and θ explicit by letting

$$l(\theta) = l(X, Y, h^{[l-1],\top}\theta) = l(X, Y, \eta),$$

where $\eta = h^{[l-1],\top}\theta$. We will abuse notation and let the derivative $l'(X, Y, \eta)$ be taken with respect to η. Since we defined $s(\theta) = \nabla_\theta l(\theta)$ in Section 2.2.2.1, $s(\theta)$ can also be written as $l'(X, Y, \eta)h^{[l-1]}$. For two functions f and g which share the same domain, we write $f \precsim g$ if $f(\cdot) \leq C \cdot g(\cdot)$. For two square matrices of the same shape, $\boldsymbol{A}_1 \prec (\preccurlyeq)\boldsymbol{A}_2$, if $A_2 - A_1$ is positive-(semi)definite. Following [**41**], we define the sub-gaussian norm ψ_2 of a random variable X as

$$||X||_{\psi_2} = \inf\left\{t > 0 : E[e^{X^2/t^2}] \leq 2\right\},$$

where $\langle ., . \rangle$ is the dot product between two vectors. The definition can be extended [**29**] to a random vector $Z \in \mathbb{R}^d$ via

$$||Z||_{\psi_2} = \sup\{||\langle Z, \theta\rangle||_{\psi_2} : ||\theta||_2 \leq 1, \theta \in \mathbb{R}^d\}. \tag{2.16}$$

For a matrix $\boldsymbol{A} \succcurlyeq 0$, we define the seminorm $||\theta||_{\boldsymbol{A}} = ||\boldsymbol{A}^{1/2}\theta||_2$ with $||\cdot||_2$ being the Euclidean norm. For a m by n matrix $\boldsymbol{A} = (a_{ij})$, we define the operator norm of $\boldsymbol{A}$ to be $\max_{x\neq 0} \frac{||\boldsymbol{A}x||_2}{||x||_2}$ and Hilbert-Schmidt norm of $\boldsymbol{A}$ to be $\left(\sum_{i,j} a_{ij}^2\right)^{1/2}$. Let $\boldsymbol{I}$ denote the identity matrix with conformable dimensions.

Now we introduced a series of assumptions from [**29**]. First, we assume the existence of matrices

$$\boldsymbol{M} = E[s(\theta_0)s(\theta_0)^\top],\ \boldsymbol{H} = E[\nabla^2 l(\theta_0)].$$

The expectations are taken with respect to the joint distribution of (X, Y, Z). H is assumed to be positive-definite.

Since our theories only cover the linear case, we may treat $h^{[l-1]}$ as the input of the linear function. To simplify the notation we are going to use Z to represent $h^{[l-1]}$–the input of the linear function J_{XY}. Next three assumptions about the distribution of Z from [**29**].

Assumption 1. *The decorrelated loss gradient at θ_0 is subgaussian.*

$$||\boldsymbol{M}^{-1/2}(s(\theta_0) - E[s(\theta_0])||_{\psi_2} \leq K_1.$$

Assumption 2. *The calibrated design $\tilde{Z} = |l''(X, Y, Z^\top\theta_0)|^{1/2}Z$ satisfies*

$$||\boldsymbol{H}^{-1/2}\tilde{Z}||_{\psi_2} \leq K_2.$$

Assumption 3. *The model is well-specified.*

$$\boldsymbol{M} = \boldsymbol{H}.$$

With Assumption 3, we fix the variance estimator to be the Fisher information when considering the score test statistic G_n under the Ising model.

Next are some probabilistic tools used in our proofs.

Theorem 2 (Theorem A.1 in [**29**]). *Let $Z \in \mathbb{R}^d$ be an isotropic (have zero mean and unit covariance) random vector with $||Z||_{\psi_2} \leq K$, and let $\boldsymbol{J} \in \mathbb{R}^{d\times d}$ be positive semidefinite. Then,*

$$P(||Z||_{\boldsymbol{J}}^2 - Tr(\boldsymbol{J}) \geq t) \leq \exp\left(-c\min\left\{\frac{t^2}{K^2||\boldsymbol{J}||_2^2}, \frac{t}{K||\boldsymbol{J}||_\infty}\right\}\right).$$

In other words, with probability at least $1-\delta$ *it holds*

$$||Z||_{\boldsymbol{J}}^2 - Tr(\boldsymbol{J}) \precsim K^2 \left(||\boldsymbol{J}||_2 \sqrt{\log(1/\delta)} + ||\boldsymbol{J}||_\infty \log(1/\delta) \right).$$

THEOREM 3 (Theorem A.2 in [**29**]). *Assume that the random vector* $X \in \mathbb{R}^d$ *satisfies* $E[\tilde{Z}\tilde{Z}^\top] = \boldsymbol{H}$ *and* $||\boldsymbol{H}^{-1/2}\tilde{Z}||_{\psi_2} \leq K$. *Let* $\boldsymbol{H}_n = \frac{1}{n}\sum_{i=1}^n \tilde{Z}_i\tilde{Z}_i^\top$, *where* $\tilde{Z}_1, \ldots, \tilde{Z}_n$ *are independent copies of* $\tilde{Z}$. *Whenever*

$$n \gtrsim K^4(d + log(1/\delta)),$$

with probability at least $1-\delta$ *it holds*

$$||\Delta||_{\boldsymbol{H}}^2/2 \leq ||\Delta||_{\boldsymbol{H}_n}^2 \leq 2||\Delta||_{\boldsymbol{H}}^2, \ \forall \Delta \in \mathbb{R}^d.$$

In other words, $H/2 \preccurlyeq H_n \preccurlyeq 2H$ *with probability at least* $1-\delta$.

COROLLARY 1. *Under the same conditions stated in Theorem 3,*

$$||\Delta||_{\boldsymbol{H^{-1}}}^2/2 \leq ||\Delta||_{\boldsymbol{H}_n^{-1}}^2 \leq 2||\Delta||_{\boldsymbol{H^{-1}}}^2, \ \forall \Delta \in \mathbb{R}^d.$$

with probability at least $1-\delta$.

PROOF. It suffices to show that for two symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \preccurlyeq \boldsymbol{B} \Rightarrow \boldsymbol{B}^{-1} \preccurlyeq \boldsymbol{A}^{-1}$. The claim requires two results. First, $\boldsymbol{A} \preccurlyeq \boldsymbol{B} \Rightarrow \boldsymbol{CAC} \preccurlyeq \boldsymbol{CBC}$ for any conformable $\boldsymbol{C}$. Second, $\boldsymbol{I} \preccurlyeq \boldsymbol{B} \Rightarrow \boldsymbol{B}$ is invertible and $\boldsymbol{B}^{-1} \preccurlyeq \boldsymbol{I}$. The first claim is straightforward to show. We prove the second claim here. Since $\boldsymbol{B}$ is symmetric, $\exists$ an orthogonal matrix $\boldsymbol{O}$ and a diagonal matrix $\boldsymbol{\Lambda}$ such that $\boldsymbol{B} = \boldsymbol{O\Lambda O}^\top$. Then $\boldsymbol{\Lambda} = \boldsymbol{O}^\top\boldsymbol{BO} \succcurlyeq \boldsymbol{O}^\top\boldsymbol{IO} = \boldsymbol{I}$ by the first result. Therefore all eigenvalues of B are larger or equal to 1. So $\boldsymbol{B}$ is invertible. Also, $\boldsymbol{B}^{-1} = \boldsymbol{B}^{-1/2}\boldsymbol{I}\boldsymbol{B}^{-1/2} \preccurlyeq \boldsymbol{B}^{-1/2}\boldsymbol{B}\boldsymbol{B}^{-1/2} = \boldsymbol{I}$. We proved the second result. Now we are ready to prove $\boldsymbol{B}^{-1} \preccurlyeq \boldsymbol{A}^{-1}$. Since

$$\begin{aligned}
\boldsymbol{A} \preccurlyeq \boldsymbol{B} &\Rightarrow 0 \preccurlyeq \boldsymbol{B} - \boldsymbol{A} \\
&\Rightarrow 0 \preccurlyeq \boldsymbol{A}^{-1/2}(\boldsymbol{B} - \boldsymbol{A})\boldsymbol{A}^{-1/2} \\
&\Rightarrow \boldsymbol{I} \preccurlyeq \boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2} \\
&\Rightarrow \boldsymbol{A}^{1/2}\boldsymbol{B}^{-1}\boldsymbol{A}^{1/2} \preccurlyeq \boldsymbol{I},
\end{aligned}$$

we have

$$\boldsymbol{B}^{-1} = \boldsymbol{A}^{-1/2}(\boldsymbol{A}^{1/2}\boldsymbol{B}^{-1}\boldsymbol{A}^{1/2})\boldsymbol{A}^{-1/2} \preccurlyeq \boldsymbol{A}^{-1/2}\boldsymbol{I}\boldsymbol{A}^{-1/2} = \boldsymbol{A}^{-1}.$$

□

LEMMA 1 (Lemma A.4 in [**29**]). *Let $Z_1, \ldots, Z_n$ be i.i.d. random vectors, then one has $||\sum_{i=1}^n Z_i||_{\psi_2}^2 \precsim \sum_{i=1}^n ||Z_i||_{\psi_2}^2$.*

LEMMA 2 (Sample mean of subgaussian is subgaussian. Lemma 5.9 in [**42**]). *Let $X_1, \ldots, X_n$ be independent mean-zero subgaussian random variables with $||X_i||_{\psi_2} = K$ for all i's. Then $\frac{1}{n}\sum_{i=1}^n X_i$ is subgaussian and $||\frac{1}{n}\sum_{i=1}^n X_i||_{\psi_2} = C_2\frac{K}{\sqrt{n}}$ for some constant C_2.*

2.6.1. Score Test Statistics. Now we are ready to present the result of the asymptotic separability of the score test statistics under the assumptions we stated. In words, asymptotic separability means as $n \to \infty$, there exists a decision rule such that the probability of making either the type-I or type-II error goes to 0. Asymptotic separability is formally stated in Corollary 2. To develop Corollary 2, we first prove Theorem 4 which says the score test statistic is bounded from above with high probability under H_0. Then in Theorem 5 we show that the score test statistic is bounded from below with high probability.

THEOREM 4 (Upper Bound of the Score Test Statistic under H_0). *Assume H_0, Assumptions 1, 2, and 3 are true. Then $G_n \precsim \frac{p(K_1^2 log(1/\delta)+1)}{n}$ with probability at least $1-2\delta$, $\delta \in (0,1)$, as long as $n \succsim p \cdot K_2^4(p + log(1/\delta))$. Recall p is the dimension of θ.*

PROOF. The proof is a slight modification of the proof of the first part of Theorem 3.1 in [**29**]. Because of Assumption 2 and the bound of n, we can apply Theorem 3 to $\boldsymbol{H}_n$ and $\boldsymbol{H}$. Therefore, we have

$$P\left(\frac{1}{2}\boldsymbol{H} \preccurlyeq \boldsymbol{H}_n \preccurlyeq 2\boldsymbol{H}\right) \geq 1-\delta.$$

Let A denote the event $\frac{1}{2}\boldsymbol{H}^{-1/2} \preccurlyeq \boldsymbol{H}_n^{-1/2} \preccurlyeq 2\boldsymbol{H}^{-1/2}$. By Corollary 1, $P(A) \geq 1-\delta$.

Next, we apply Assumption 1 to prove the claim. Under H_0, ∇s_i are independent, zero mean, and with covariance $\boldsymbol{M}$. Hence, random vectors $\boldsymbol{M}^{-1/2}s_i(\theta_0)$ are independent and isotropic. By Assumption 1 and $E[s_i(\theta_0] = 0$ under H_0, $||\boldsymbol{M}^{-1/2}s_i(\theta_0)||_{\psi_2} \leq K_1$ for all i's. By Lemma 1

about the subgaussian norm of the sum of i.i.d. random vectors, we have that the random vector $\boldsymbol{A}_n = \sqrt{n}\frac{1}{n}\sum_{i=1}^n \boldsymbol{M}^{-1/2} s_i(\theta_0)$ is isotropic, satisfying $||\boldsymbol{A}_n||_{\psi_2} \precsim K_1$. Moreover,

$$||\frac{1}{n}\sum_{i=1}^n s_i(\theta_0)||^2_{\boldsymbol{H}^{-1}} = \frac{1}{n}||\boldsymbol{A}_n||^2_{\boldsymbol{J}}, \text{ with } \boldsymbol{J} = \boldsymbol{M}^{1/2}\boldsymbol{H}^{-1}\boldsymbol{M}^{1/2}.$$

By Theorem 2, Assumption 3, and the fact that $||\boldsymbol{J}||_\infty \leq ||\boldsymbol{J}||_2 \leq Tr(\boldsymbol{J}) = p$,

$$P(||\boldsymbol{A}_n||^2_{\boldsymbol{J}} \precsim p(K_1^2\log(1/\delta)+1)) \geq 1-\delta \Rightarrow P(||\frac{1}{n}\sum_{i=1}^n s_i(\theta_0)||^2_{\boldsymbol{H}^{-1}} \precsim \frac{p(K_1^2\log(1/\delta)+1)}{n}) \geq 1-\delta$$

Let B denote the event $||\frac{1}{n}\sum_{i=1}^n s_i(\theta_0)||^2_{\boldsymbol{H}^{-1}} \precsim \frac{p(K_1^2 log(1/\delta)+1)}{n}$. Then

$$\begin{aligned}
P(G_n \precsim \frac{p(K_1^2\log(1/\delta)+1)}{n}) &= P(\left\|\frac{1}{n}\sum_{i=1}^n s_i(\theta_0)\right\|^2_{\boldsymbol{H}_n^{-1}} \precsim \frac{p(K_1^2\log(1/\delta)+1)}{n}) \\
&\geq P(\left\|\frac{1}{n}\sum_{i=1}^n s_i(\theta_0)\right\|^2_{\boldsymbol{H}_n^{-1}} \precsim \frac{p(K_1^2\log(1/\delta)+1)}{n} \cap A \cap B) \\
&= P(A\cap B) \\
&\geq P(A)+P(B)-1 \\
&\geq 1-2\delta.
\end{aligned}$$

□

THEOREM 5 (Lower Bound of the Score Test Statistic under H_1). *Assume H_1, Assumption 1, 3 and $n \succsim p\cdot K_2^4(p+log(1/\delta))$. With probability $1-3\delta$,*

$$G_n \geq E\left[\left\|\frac{1}{n}\sum_{i=1}^n s_i(\theta_0) - E[s(\theta_0)]\right\|^2_{\boldsymbol{H}^{-1}}\right] - \min\{t_1,t_2\} - t_3 + ||E[s(\theta_0)]||^2_{\boldsymbol{H}^{-1}},$$

where $t_1 = (\frac{1}{c}\log\frac{2}{\delta})^{\frac{1}{2}} C_2^2\frac{K_1^2}{n}p$, $t_2 = \frac{C_2^2}{c}\log\frac{2}{\delta}\frac{K_1^2}{n}p$, and $t_3 = \left(\frac{C_2}{c}\log\frac{2}{\delta}\frac{K_1^2}{n}\right)^{1/2}||E[s(\theta_0)]||_{\boldsymbol{H}^{-1}}$ for some constant c and C_2.

PROOF.

$$||\frac{1}{n}\sum_{i=1}^{n} s_i(\theta_0)||^2_{\boldsymbol{H}^{-1}} = \underbrace{\left\|\frac{1}{n}\sum_{i=1}^{n} s_i(\theta_0) - E[s(\theta_0)]\right\|^2_{\boldsymbol{H}^{-1}}}_{a} + \underbrace{2\,(\frac{1}{n}\sum_{i=1}^{n} s_i(\theta_0) - E[s(\theta_0)])^\top \boldsymbol{H}^{-1} E[s(\theta_0)]}_{b} + ||E[s(\theta_0)]||^2_{\boldsymbol{H}^{-1}}$$

Let $\boldsymbol{J} = \boldsymbol{M}^{1/2}\boldsymbol{H}^{-1}\boldsymbol{M}^{1/2}$, then $a = ||\frac{1}{n}\sum_{i=1}^{n} G^{-1/2}\,(s_i(\theta_0) - E[s(\theta_0)])\,||^2_{\boldsymbol{J}}$. By Assumption 1, $\forall i \in \{1,\ldots,n\}, \boldsymbol{M}^{-1/2}\,(s_i(\theta_0) - E[s(\theta_0)])$ is subgaussian with parameter K_1. Then by Lemma 2, $\sum_{i=1}^{n} \frac{1}{n}\boldsymbol{M}^{-1/2}\,(s_i(\theta_0) - E[s(\theta_0)])$ is also a subgaussian with parameter $C_2\frac{K_1}{\sqrt{n}}$ for some constant C_2. By Hanson-Wright inequality [**34**], for any $\delta \in (0,1)$

$$P(|a - E[a]| > \min\{t_1, t_2\}) > 1 - \delta$$

$$\Rightarrow P(a > E[a] - \min\{t_1, t_2\}) > 1 - \delta$$

where $t_1 = (\frac{C_2^4}{c}\log\frac{2}{\delta}\frac{K_1^4}{n^2}||\boldsymbol{J}||^2_{HS})^{\frac{1}{2}}$, $t_2 = \frac{C_2^2}{c}\log\frac{2}{\delta}\frac{K_1^2}{n}||\boldsymbol{J}||$ and c is a constant. With Assumption 3, $J = \boldsymbol{I}$ and $t_1 = (\frac{1}{c}\log\frac{2}{\delta})^{\frac{1}{2}}C_2^2\frac{K_1^2}{n}p$, $t_2 = \frac{C_2^2}{c}\log\frac{2}{\delta}\frac{K_1^2}{n}p$. Now we bound term b.

$$\begin{aligned} b &= E[s(\theta_0)]^\top \boldsymbol{H}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} s_i(\theta_0) - E[s(\theta_0)]\right) \\ &= E[s(\theta_0)]^\top \boldsymbol{H}^{-1}\boldsymbol{M}^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{M}^{-1/2}\,(s_i(\theta_0) - E[s(\theta_0)])\right] \end{aligned}$$

Let $g = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{M}^{-1/2}\,(s_i(\theta_0) - E[s(\theta_0)])$. Since g is a subgaussian random vector, each entry g_i in the random vector is a subgaussian random variable with $||g_i||_{\psi_2} \le ||g||_{\psi_2} = C_2\frac{K_1}{\sqrt{n}}$. This can be seen by setting θ in (2.16) to be one-hot vectors. By General Hoeffding Inequality (Theorem 2.6.3 in [**41**]), we have

$$P(|b| < t_3) > 1 - \delta$$

$$\Rightarrow P(b > -t_3) > 1 - \delta$$

where $t_3 = \left(\frac{C_2}{c}\log\frac{2}{\delta}\frac{K_1^2}{n}\right)^{1/2}||E[s(\theta_0)]^\top \boldsymbol{H}^{-1}\boldsymbol{M}^{1/2}||_2$. Apply Assumption 3,

$$t_3 = \left(\frac{C_2}{c}\log\frac{2}{\delta}\frac{K_1^2}{n}\right)^{1/2}||E[s(\theta_0)]^\top \boldsymbol{H}^{-1/2}||_2 = \left(\frac{C_2}{c}\log\frac{2}{\delta}\frac{K_1^2}{n}\right)^{1/2}||E[s(\theta_0)]||_{\boldsymbol{H}^{-1}}$$

So with probability at least $1-2\delta$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n} s_i(\theta_0)\right\|^2_{\boldsymbol{H}^{-1}} > E[a] - \min\{t_1, t_2\} - t_3 + ||E[s(\theta_0)]||^2_{\boldsymbol{H}^{-1}}.$$

Let A denote the event $\frac{1}{2}\boldsymbol{H}^{-1/2} \preccurlyeq \boldsymbol{H}_n^{-1/2} \preccurlyeq 2\boldsymbol{H}^{-1/2}$. Since, $n \gtrsim p \cdot K_2^4(d + log(1/\delta))$, by Corollary 1, $P(A) \geq 1-\delta$. Using similar argument in Theorem 4, we conclude that with probability at least $1-3\delta$,

$$G_n = \left\|\frac{1}{n}\sum_{i=1}^{n} s_i(\theta_0)\right\|^2_{\boldsymbol{H_n}^{-1}} > E[a] - \min\{t_1, t_2\} - t_3 + ||E[s(\theta_0)]||^2_{\boldsymbol{H}^{-1}}.$$

□

Corollary 2 (Asymptotic Separability). *Under Assumption 1, 2, 3, there exists a sequence $t_n \in (0, \infty)$, such that the decision rule rejecting H_0 if $G_n > t_n$ has*

$$P(\textit{Reject } H_0|H_0) + P(\textit{Fail to reject } H_1|H_1) \to 0,$$

as $n \to \infty$.

Proof. Let $a = \frac{p(K_1^2 log(1/\delta)+1)}{n}$ be the bound in Theorem 4 and

$$b = E\left[\left\|\frac{1}{n}\sum_{i=1}^{n} s_i(\theta_0) - E[s(\theta_0)]\right\|^2_{\boldsymbol{H}^{-1}}\right] - \min\{t_1, t_2\} - t_3 + ||E[s(\theta_0)]||^2_{\boldsymbol{H}^{-1}}$$

be the bound in Theorem 5. It is suffice to find a t_n such that $a < t_n < b$ for large enough n. As $n \to \infty$, a, t_1, t_2, and t_3 converges to 0. $E\left[\left\|\frac{1}{n}\sum_{i=1}^{n} s_i(\theta_0) - E[s(\theta_0)]\right\|^2_{\boldsymbol{H}^{-1}}\right]$ is non-negative as H is assumed to be positive-definite in this section. So as $n \to \infty$, $a \to 0$ and

$$b - E\left[\left\|\frac{1}{n}\sum_{i=1}^{n} s_i(\theta_0) - E[s(\theta_0)]\right\|^2_{\boldsymbol{H}^{-1}}\right] \to ||E[s(\theta_0)]||^2_{\boldsymbol{H}^{-1}}.$$

Thus,

$$\lim_{n\to\infty} b \geq ||E[s(\theta_0)]||^2_{\boldsymbol{H}^{-1}}$$

Hence, $\exists M > 0, t_n \in (0,1)$ such that $n > M \Rightarrow a < t_n < b$. □

Now, we have shown the asymptotic separability of the score test statistics under the Ising model assuming linear $J(Z)$. Although our theories are developed under strong assumptions, we have demonstrated that our methods have valid type-I error control and power in both simulations and two real-world applications (the adult data set and COOS data). Our future work will focus on two areas. First, we would like to extend the separability to more general cases (e.g. non-linear $J(Z)$ and the RBM model). Next, we want to show that the distribution of the bootstrap test statistic produced by the score based bootstrap procedure is close to the true distribution of the test statistic.

CHAPTER 3

Multiple Imputation for Detecting Covid-19 via Wastewater-based Epidemiology

3.1. Introduction

The project is a collaboration with the Wastewater team through the Healthy Davis Together program in Davis. For a more detailed discussion of motivation, methods, and results of the overall wastewater monitoring program, please refer to Hannah et al.'s paper [**36**]. We will provide a summary of the background of the project described in [**36**], then we will dive into the algorithm.

Wastewater-based epidemiology (WBE) is often used as a supplement to clinical data. Researchers hope to use the signal from wastewater data as a leading indicator of a potential outbreak. WBE doesn't require the public to change their behaviors (e.g. require residents to participate in COVID testing). Therefore, not only WEB is often a cheaper alternative to mass testing, but also it reduces the sampling bias as the data is generated from the entire population instead of people who show up for testing or treatment.

HDT ran wastewater surveillance in Davis by collecting wastewater samples from maintenance holes across various locations. "Sample extracts were analyzed by one-step RT-qPCR for four targets: N1 and N2 targeting regions of the nucleocapsid (N) gene of SARS-CoV-2, ϕ6 bacteriophage (an RNA virus used as an internal quality control), and pepper mild mottle virus (PMMoV; used for normalization of SARS-CoV-2 results)". [**36**] Measuring the N1 and N2 genes alone is not enough, because the virus concentration level can change with based on the number of people who produce the wastewater. PMMoV is used for normalization because it is a common RNA virus found in human feces. Figure 3.4 shows the variability of the PMMoV despite being a common virus produced by humans all the time. Hence, normalization is required to correctly measure the COVID-19 gene concentration level relative to the population in the study. For the

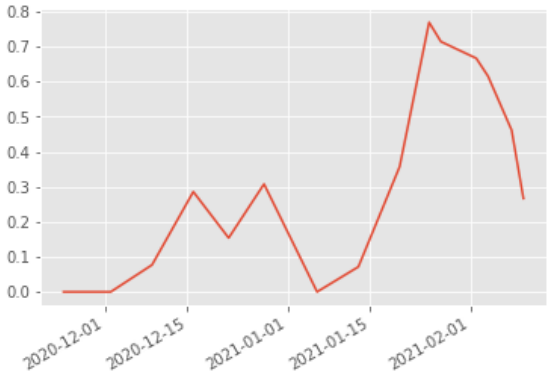


FIGURE 3.1. Proportion of Nondetects Overtime

imputation task, N1 and N2 targets are what our project focused on. Roughly speaking, the RT-qPCR procedure produces a Ct value for the sample collected at each location. The Ct value indicates the concentration level of the gene of interest. The lower the number, the higher the concentration is. The Ct value is truncated from above at 45 and naturally bounded below at 0. We will conclude that the target is not detected at the sample if the Ct value is at the upper boundary.

For these non-detect samples, the wastewater team requested an imputation method to impute qPCR values. Our task was to help the wastewater team to develop an imputation model and provide an easy-to-use API.

3.2. Nondetects in qPCR Data

Why are nondetects problems in analyzing qPCR data? Nondetects happen on regular basis but not at random. From Figure 3.3 we don't see a clear trend of the detected Ct value of the target gene at each sub-location. However, Figure 3.1 clearly shows that the occurrence nondetects of target genes have patterns over time with a large peak in late January. This should correspond with a decrease in infections. So nondetects contain a lot of information and improper handling may introduce bias to the analysis. A nondetect could represent one of the several possibilities: (1) low starting target abundance, (2) complete absence of a target from the sample, or (3) human error / experimental failure [**36**]. Existing methods typically use a single value to impute nondetects. For possibility (1), using a high Ct value to impute may be reasonable, whereas, using 0 may be

reasonable for possibility (2). Dropping nondetects is appropriate for possibility (3). In reality, nondetects are mostly likely to be a mixture of these three possibilities. Therefore, the single value imputation method may lead to bias in the analysis. A more sophisticated imputation method is required to produce a valid analysis.

3.3. Model and Algorithm

We model the Ct value at each maintenance hole using a Bayesian model. Let Y_{ij} denote the Ct value at location i and jth replicates. Each location has three readings captured by different techniques [**36**], so $j \in \{1, 2, 3\}$. Then we assume

$$Y_{ij} \overset{\text{i.i.d.}}{\sim} \text{truncated normal}(\theta_i, \sigma^2, 0, \infty)$$

$$\theta_i \sim \text{Gamma}(\alpha_\theta, \beta_\theta)$$

$$\sigma \sim \text{Gamma}(\alpha_\sigma, \beta_\sigma)$$

Our exploratory analysis provides some justification for our model. Figure 3.2 shows that the log(N1 c/L) variable can be reasonably said to be normally distributed, with a standard deviation of 4.13. From Figure 3.3, we can also see that there are no strong trends for the N1 c/L variable over the location and each location seem to have different means. Hence, we model Y_{ij} with a truncated normal and each location has its own mean θ_i.

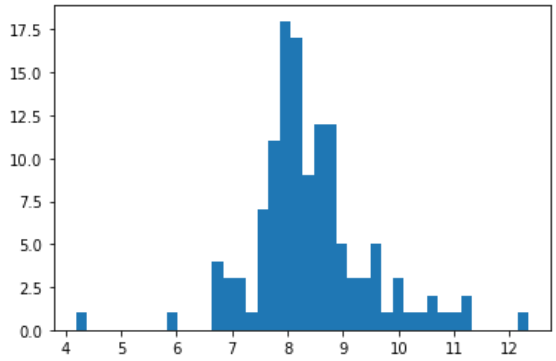


FIGURE 3.2. Distribution of the N1 Target Gene. The unit is log gene copies per liter (gc/L) log-scaled.

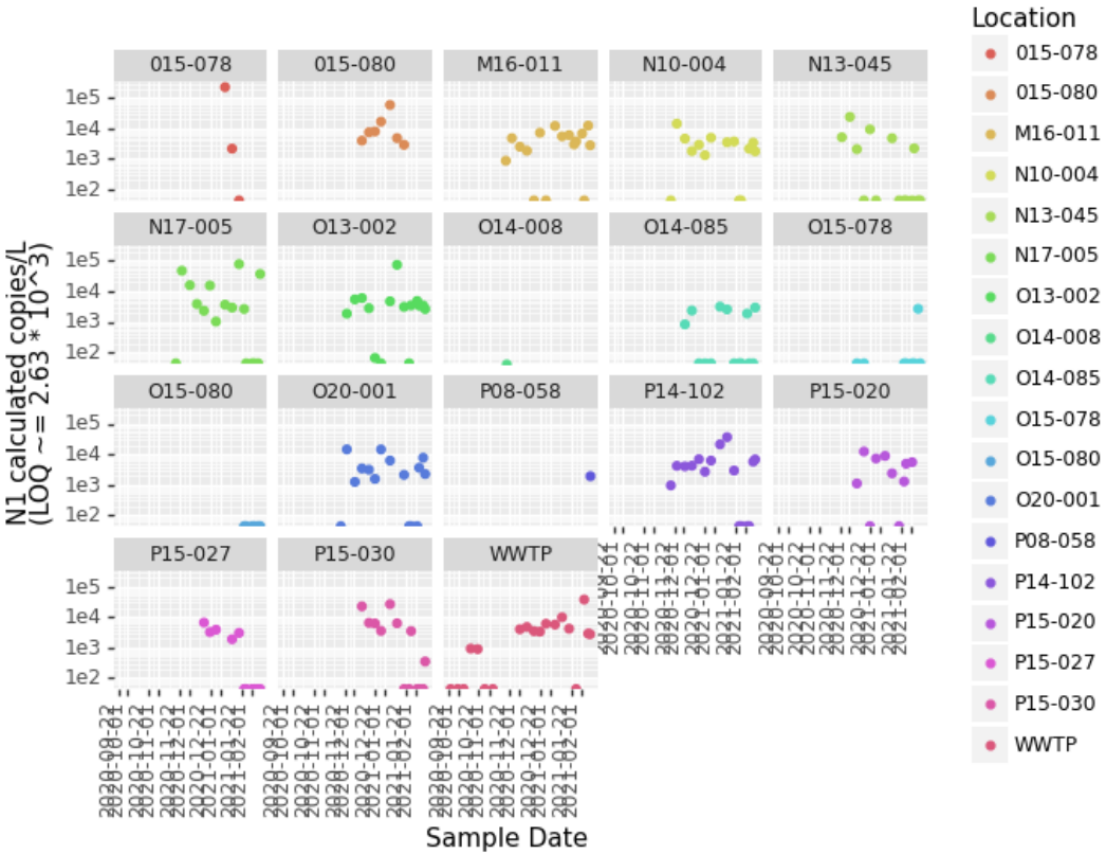


FIGURE 3.3. N1 (gc/L) by location log-scaled. In this figure, we impute the measurements below LOD with 0. These are represented as points at the lower bound of the plot.

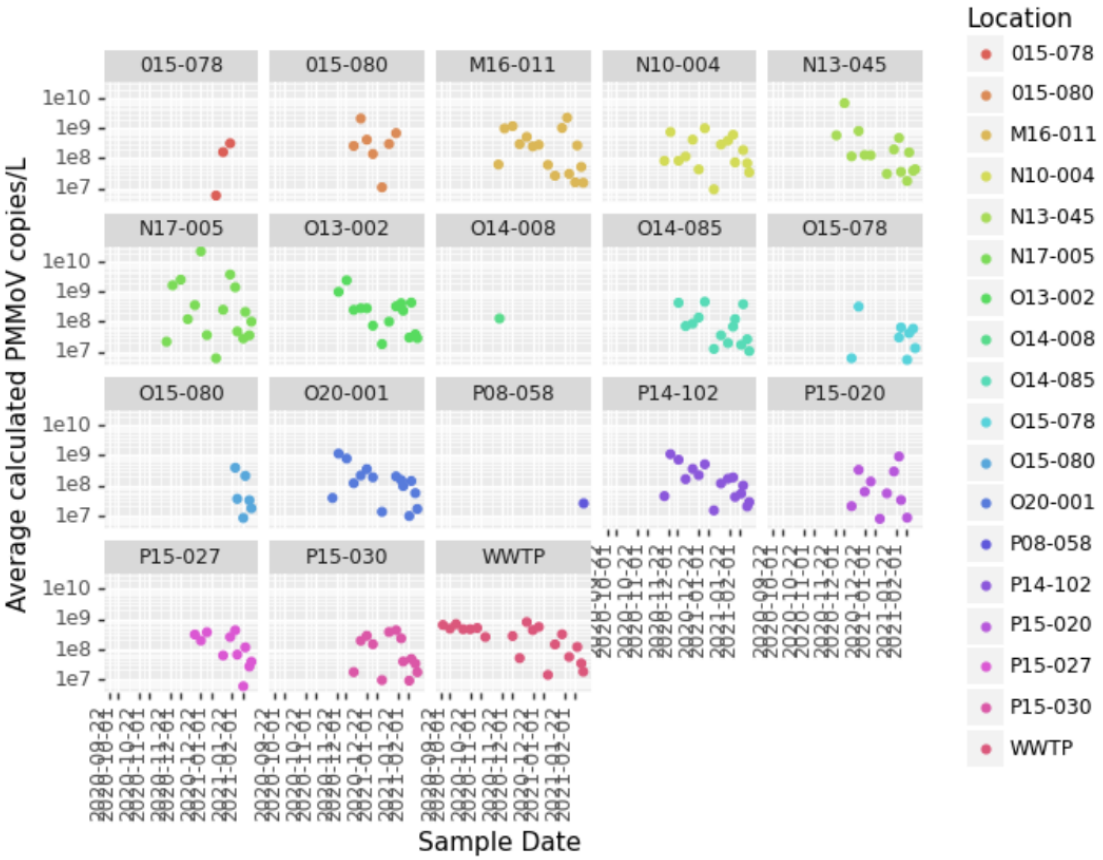


FIGURE 3.4. PMMOV (gc/L) by location log-scaled. All measurements of PMMOV are above the LOD.

To simulate from the posterior distribution of $\theta_i|Y_i$, we apply the EM-MCMC algorithm (also known as the empirical Bayes [**7**]) via PyStan [**32**].

Algorithm 3 EM-MCMC

Input: $\{Y_{ij}\}_{i,j}^n$, $\alpha_\theta, \beta_\theta, \alpha_\sigma, \beta_\sigma$, iteration ($N$), burn-in sample ($B$), number of MCMC samples (T).

for $b \in \{1 \ldots, N\}$ **do**

For all i's:

Produce the posterior distribution $P(\theta_i|Y_i)$.

Generate T MCMC $\tilde{\theta}_i$ according to $P(\theta_i|Y_i)$ and drop the first B of them.

Update α_θ and β_θ with the maximum likelihood estimation based on all $\tilde{\theta}_i$'s across all i's.

end for

Return: $P(\theta_i|Y_i)$ for all i's.

B should large enough so that until $\alpha_\theta, \beta_\theta$ converges. Then we use the posterior distribution to compute the posterior mean of θ_i for all i's. We provide a user-friendly API so that users with minimum Python experience can apply Algorithm 3 to real-world data with ease. As demonstrated in Hannah et al.'s source code [**36**], our API allows users to run the Algorithm without having to write in Stan modeling language which is difficult to debug. Before shipping the Algorithm and Python Script to our collaborator, we provide an example by carrying out Algorithm 3 to the Wastewater data collected on Jan 7th, 2021 to determine whether the method is able to produce sensible inference results. We initialize $\alpha_\theta = 1, \beta_\theta = \frac{1}{35}, \alpha_\sigma = 3, \beta_\sigma = 1$ and set $N = 15$, $B = 500$, $T = 10^4$.

We first examine the trace plots of parameters from the last iteration in Algorithm 3 through Figure 3.5.

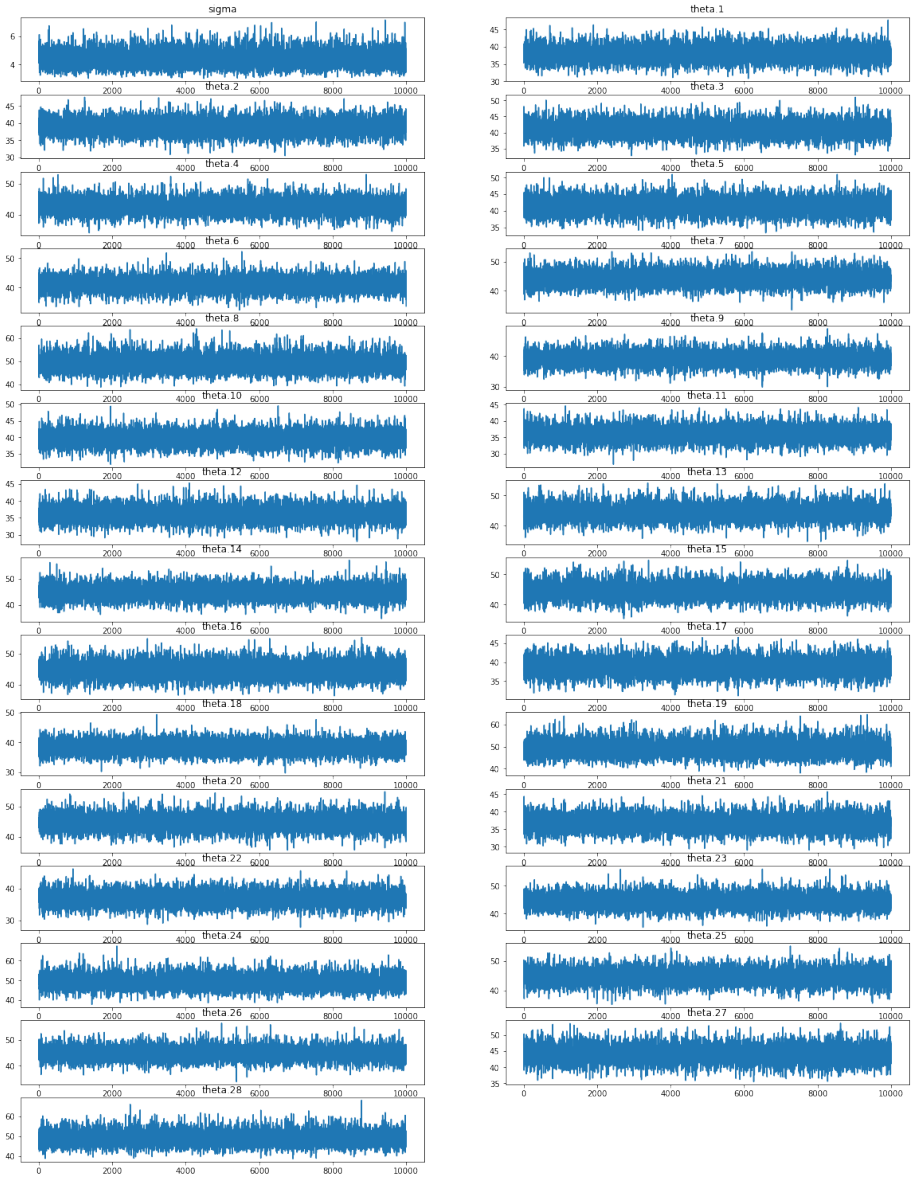


FIGURE 3.5. Trace plots of parameters during one iteration. The Markov chains didn't stick in a local region.

Since the trace plots don't show discernible patterns, Markov chains in the last iteration are able to reach the stationary distribution quickly. Furthermore, Figure 3.6 and 3.7 show that both hyper-parameters $\alpha_\theta, \beta_\theta$ and the posterior means of parameters are able to converge on the selected data file. $\alpha_\theta, \beta_\theta$ converges to 80.98 and 7.91, so the estimated prior distribution of θ has mean 42.4 and standard deviation 4.71. Given the context, the estimated prior distribution is reasonable.

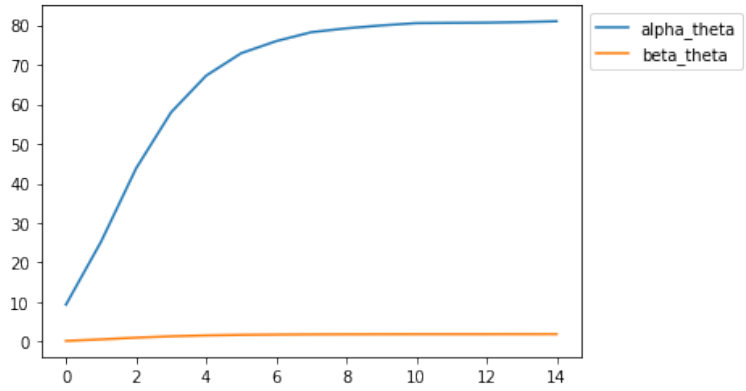


FIGURE 3.6. The graph shows the trace $\alpha_\theta, \beta_\theta$ over 15 iterations. The horizontal axis shows the iteration and the vertical axis shows the parameter value.

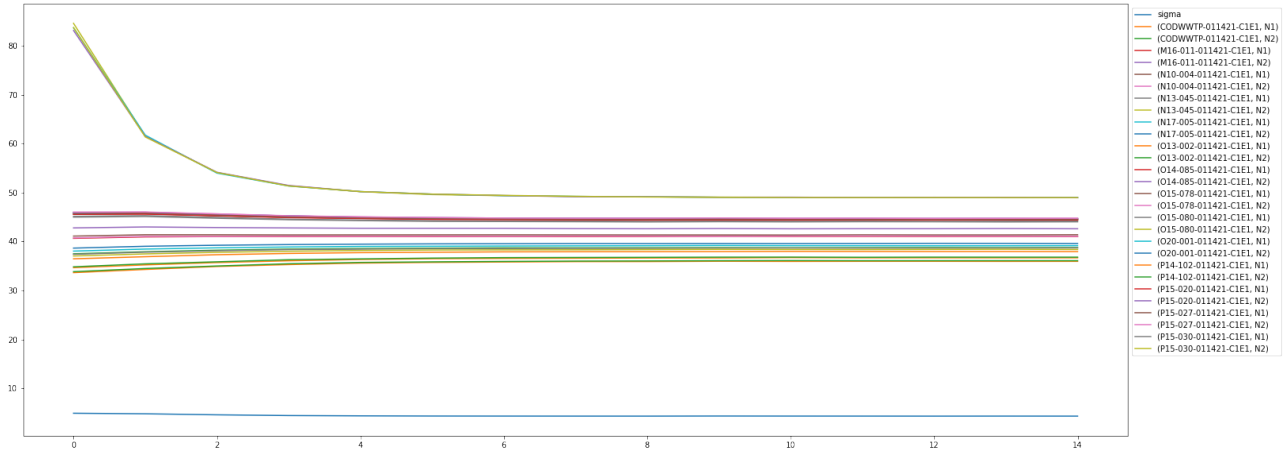


FIGURE 3.7. The graph shows the trace σ and θ_i's over 15 iterations. The horizontal axis shows the iteration and the vertical axis shows the parameter value.

Most posterior means of θ_i's are around 40 which is close to the estimated prior mean. Those locations with high posterior means of θ_i's all have observed Ct values truncated at 45, therefore we expect to see large posterior means. Finally, the posterior mean of σ is 4.37. So our EM-MCMC algorithm is able to provide plausible inference results to be used for modeling and imputation on the full data. In the next section, we will show the results of the EM-MCMC algorithm and compare it to three other methods.

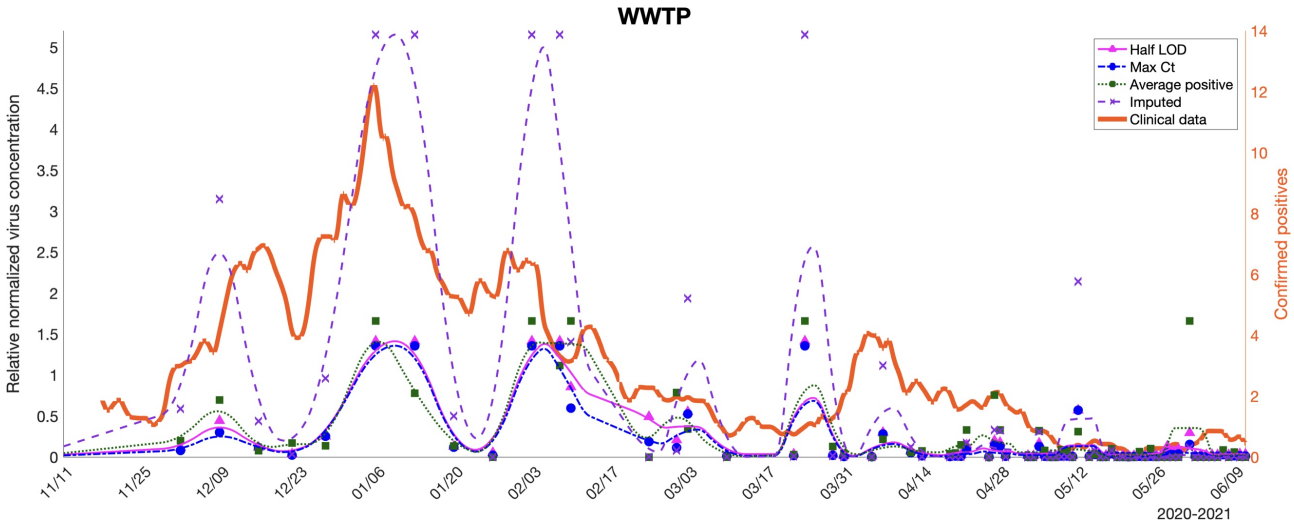


FIGURE 3.8. Community-level wastewater vs clinical data in Davis, showing effects of different methods of handling nondetects. Symbols represent individual sample results; lines represent trends (as centered 7-day moving averages for all data shown).

3.4. Result

We include three other imputation methods for comparison: $\text{LOD}_{0.5}$ (single imputation with half the detection limit), Ct_{max} (single imputation with the maximum qPCR cycle) and Ct_{avg} (dropping all nondetects from the analysis). Figure 3.8 shows the normalized COVID-19 gene concentration level with the clinical data. Different imputation methods produce similar overall trends, but there are still differences. Compared to other methods, $\text{LOD}_{0.5}$ produced higher virus concentration when close to the end of the experiment period. On some dates, one method produced much higher concentration values compared to other methods. For example, the data produced by EM-MCMC is much higher for December 9. To provide a more quantitative assessment of these imputation methods, we compute Spearman's rank-order correlation between the clinical data and data imputed. The correlation analysis has its limitations. Hannah et al. [**36**] state that the clinical data may lag the trend from the wastewater data despite the low probability of a systematic lag. Table 3.1 shows that overall values produced by the EM-MCMC algorithm have the strongest correlation with the clinical data. So the empirical results suggest that the EM-MCMC algorithm yields better results than more common and nondetect handling methods.

$LOD_{0.5}$	Ct_{max}	Ct_{avg}	EM-MCMC
0.2175	0.5049	0.4337	0.5447

TABLE 3.1. Spearman's Rank-Order Correlation Coefficients between Community-Level Clinical Cases and Relative Normalized Virus Concentration produced by Nondetect Handling Method.

CHAPTER 4

Assessing the Impact of College Reopening on Covid-19 Outbreaks

4.1. Introduction

Studies suggest that during the period of college reopening, confirmed cases in college-towns increased overwhelmingly [**18**, **25**]. Figure 4.1 compares COVID-19 case and death per 100,000 individuals as reported by JHU CSSE [**10**] between counties with a large undergraduate population to the overall US counts[1]. In Section 4.2, we describe our classification of "college counties", namely counties with an undergraduate population of all colleges in the county exceeds 10% of the population. One can naively assume that this means that the effective reproduction number (R_T)—the expected number of secondary infections that a single infection will generate—was extremely high when students returned to their universities for the fall semester. [**25**] estimates that one such college experienced a maximum R_T value of 10.75 (University of Washington), and extremely high number relative to prior estimates at the national and sub-national scale [**1**]. These estimates were fit using an SEIR model, that does not account for the importation of cases, as well as biases in reporting due to testing regime.

We also observed a significant rise in cases within college counties, however a few additional observations stand out. First, through the remainder of the fall term, the case counts decrease and roughly track with the US average. This includes the time period through and following the winter break of 2020-2021. Second, the death rate in these counties do not increase dramatically in concert with the confirmed cases. While the death rates do approach the US average during the reopening period, overall college counties had lower deaths per capita than the US on average average.

[1]We access the number of daily COVID-19 cases and deaths between May 1st, 2020, and February 21st, 2021 through the COVIDcast Epidata API [**14**]. The API is built and maintained by the Carnegie Mellon University Delphi research group.

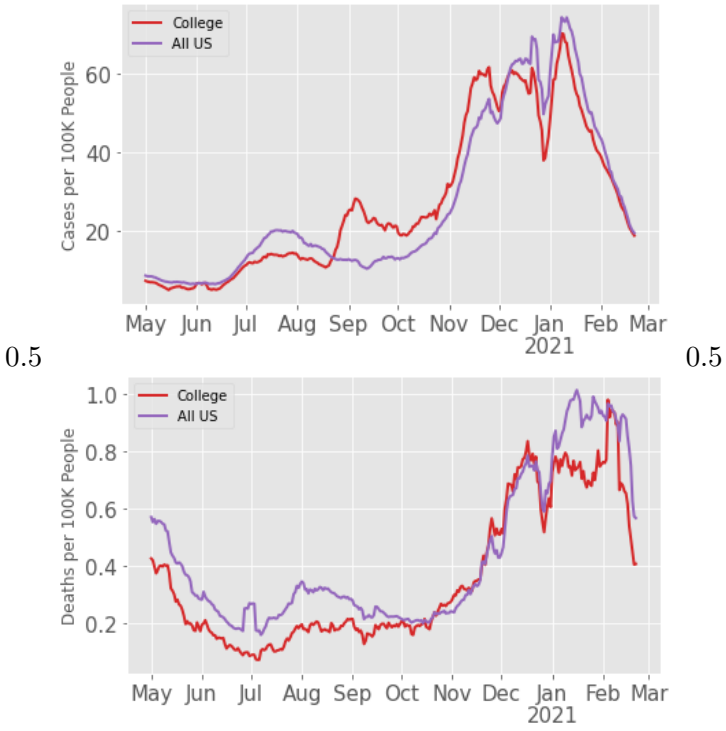


FIGURE 4.1. 7-Day moving average of cases per 100,000 people (left) and 7-Day moving average death per 100k people (right). Data reported by JHU CSSE [**10**].

Should we blame this increase in cases on students returning to campus and engaging in behavior that promoted the spread of SARS-CoV-2. In fact, assuming that the increase in cases is simply due to increased transmission among college students neglects other possible factors. First, the college counties may not mirror the US average because of their locations and demographics. If we were to compare the counties to similar non-college counties then we may see similar case trajectories. We examine this problem by matching the college counties to non-college counties that are within the same state and have similar percentage of seniors (an important demographic variable for COVID-19 rates). Second, could the increase in cases actually be spurious, and this observation is simply a false alarm due to noisy case reporting? We cast this question as a multiple hypothesis test and define a notion of hotspots that controls the false discovery rate. Third, are these hotspots associated with how the colleges reopened, such as in-person classes or the testing availability on campus? We investigate in greater detail a selection of colleges within the college

counties and categorize them in terms of their COVID-19 mitigation measures. Then we test for associations between the measures and hotspot status.

4.1.1. Summary of Results. We summarize the findings of our analysis.

(1) Colleges counties experienced a higher than expected proportion of outbreaks around the start of 2020 Fall instruction relative to their matched non-college counties.

(2) Class setting (in-person, hybrid, etc.) may affect the severity of the outbreak at college counties (p-value: 0.0157); however, with the Bonferroni multiple testing correction we do not achieve a 5% significance level at the corrected threshold of 0.01.

(3) Testing availability of colleges is not significantly associated with hotspot status (p-value: 0.0634), however with sufficient data this association could be clarified.

(4) After the initial outbreak, the proportion of college counties which were hotspots were not higher than non-college counties. Colleges were not a major factor driving up the infection level after the beginning of the Fall instruction.

4.2. Matched County Analysis

In order to control for the possible demographic and region effects on the case and death incidences, we match college counties to similar counties based on demographics and location.

First, we must define in greater detail what constitutes a "college county". A county is considered as a college county if at least 10% of its population are undergraduate students. We use US Census Bureau data [**5**] to estimate the county population and the College Scorecard dataset to estimate the county undergraduate population. The College Scorecard data is maintained by the US Department of Education, and it is an institution level summary of college metrics for all institutions that receive federal financial aid. Included in the dataset are the counties in which the campus resides, the undergraduate student population, and the degrees conferred by the colleges among many other variables. Our analysis only included four-year institutions and left out institutions such as 2-year vocational schools. Based on our definition, 152 counties are college counties [2].

[2]Initially, there were 153 counties satisfying our criterion. However, we manually exclude the Salt Lake County. See Appendix 4.6.1 for details.

One might expect that college counties have unique demographic characteristics. In order to control for demographic factors, we find their population density and proportion of seniors (aged 65 and above). We found that the senior population was sufficient for the matching process, and matching based on more demographic bins did not significantly impact the matched counties.

	College County			Matched Non-College County		
State	County	Pop Density	65+ %	County	Pop Density	65+ %
CA	Yolo	217.3	13.7	Fresno	167.7	13.6
NY	Albany	584.4	16.8	Onondaga	591.6	17.1
OH	Athens	129.7	13.6	Union	136.6	13.9
TX	Brazos	391.5	11.1	Webb	82.3	11.4
UT	Cache	110.1	11.4	Tooele	10.4	11.1

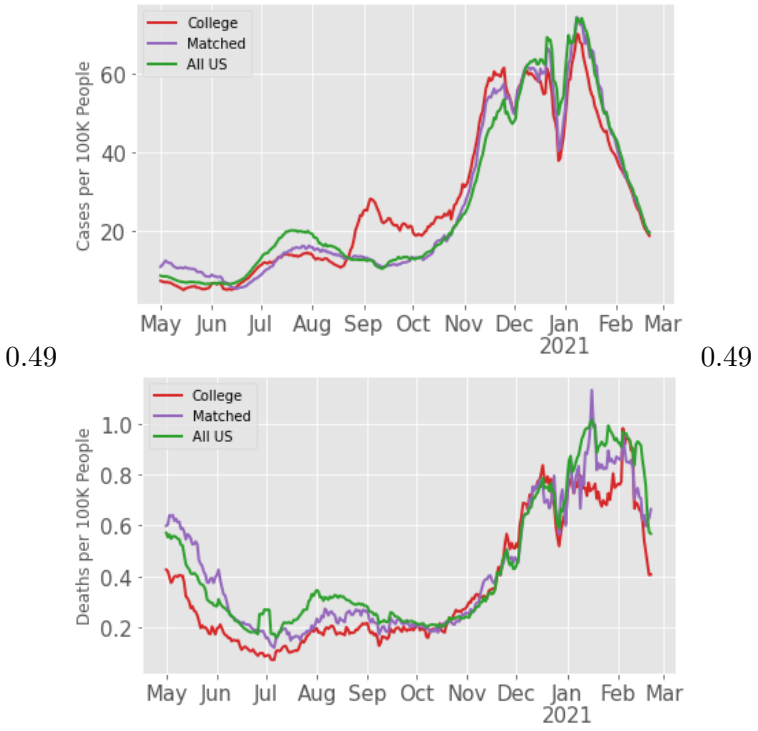
TABLE 4.1. Examples of Matched Non-College Counties.

4.2.0.1. *Matching Process.* We use the Hungarian algorithm [**33**], to optimally assign the matched pairs of college and non-college counties within the same state based on the Euclidean distances between the standardized variables (senior proportion and population density). See Appendix 4.6.2 for more details. Table 4.1 contains five matching results and the algorithm behaved as expected. The pair, Albany county (college) and Onondaga county (non-college), in New York state is an example of similar population densities (584.365 vs. 591.642) and senior proportions (16.838% vs. 17.146%). However, it is often impossible to find a matching non-college county with a similar population density in many states, and the algorithm will primarily match based on the senior proportion. For example, the population density of Cache county in Utah is more than ten times higher than Tooele County (110.137 vs. 10.41). However, they were matched because of their similar 65 plus percentages (11.361% vs. 11.139%).

	College Counties	Matched Counties	All US Counties
Cases	8133.2	7931.4	8001.7
Deaths	106.1	124.2	129.6

TABLE 4.2. Total 7-Day Average Cases & Deaths per 100k People between 1st May, 2020 and 21st Feb, 2021.

4.2.0.2. *Paired Analysis.* We compare college counties and non-college counties in terms of COVID-19 reported case and death incidences. Table 4.2 shows that matched non-college counties



0.49 0.49

FIGURE 4.2. Comparison of college counties to their matched counterparts and the US total for case counts (left) and death counts (right). All results are 7 day averages of incidences per 100K people.

had similar case and death incidence per 100k people compared with all US counties (over the entire study period). College counties had slightly more cases but notable fewer deaths on average compared to matched counties. In Figure 4.2 (left), we see that the 7-day average cases per 100k people of matched non-college counties tracked closely with all US counties. COVID-19 cases of non-college and all US counties were gradually decreasing between mid-August to October when most US colleges began their Fall 2020 instructions. On the other hand, 7-day average cases per 100k people of college counties were similar to matched non-college counties but diverged between mid-August and October. Over the one and a half months, college counties were experiencing a clear spike in COVID-19 cases relative to their matched counterparts. The biggest difference (15.665) between 7-day average cases per 100k people of college and non-college counties occurred on 6 Sept. 2020.

In Figure 4.2 (right), the curve of death per 100k people in college counties is almost indistinguishable from the matched counties between September 2020 and January 2021. [**25**] suggests

that colleges were spreading the virus to their home counties and [**18**] indicates that the quickly rising death in college towns was linked to colleges. But if either claim were accurate, we would have observed that deaths per 100k, as the case per 100k, at college counties to be significantly higher than the matched counties after the reopening of colleges. Recall that more often than not, our matching algorithm matched non-college counties based on the percentage of senior people who are a vulnerable group to COVID-19. Therefore, the low total death count in Table 4.2 and the similarity of the death rate trajectories suggests that rising cases were primarily spread among young college students instead of local communities.

4.3. Hotspot Identification

The purpose of our hotspot analysis is to test whether this increase can be attributed to noise in the data. We focus on an increase in cases as opposed to simply a large case count because the increase is more closely related to the effective reproduction number (R_T). To this end we define a hotspot and compare the hotspot incidences between college counties and their matches.

4.3.1. Hotspot definition. How should we define hotspots? Simply observing an increase in cases does not necessarily imply that a county is currently a hotspot. Our definition controls for false alarms by casting the problem as a multiple hypothesis test, and it requires that the increase be sufficiently large (exceeding a 10% increase between two sequential 14 day time windows). This definition gives us added certainty that the increase in cases is not due to randomness in the reported cases, but is a real increase in the unobserved Covid-19 incidence rate. An added bonus of using hotspots is that the hotspot definition is reflective of the relative increase in cases as opposed to the magnitude of the case counts, which is more reflective of R_T.

Specifically, we compare the county case counts 14 days before and 14 days after a pivot date over a grid of pivot dates. For each pivot date, county pair we postulate a null hypothesis in which the expected counts (following a Poisson model) does not increase by greater than 10%. We then perform the Benjamini-Hochberg (BH) procedure [**2**] for multiple hypothesis testing which controls the false discovery rate (FDR) at a 5% level. The FDR is defined to be the percentage of the true null hypotheses that are falsely rejected. The BH procedure is guaranteed to control FDR at level α assuming that the tests are independent, although there are weaker conditions in which FDR is

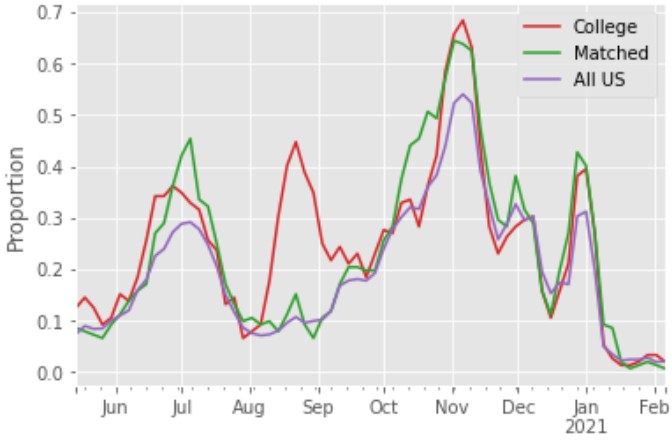


FIGURE 4.3. Proportion of Hotspots

FIGURE 4.4. Proportion of counties that are labeled hotspots for college counties, their matched counterparts, and the US total.

controlled [**3**]. Finally, all of the county, date pairs that correspond to a rejected null hypothesis is considered to be a *hotspot*. See appendix 4.6.3 for more details.

4.3.2. Hotspot Analysis. Comparing Fig. 4.2 (left) and Fig. 4.4, we see that our hotspots labeling method is a reasonable automated way to label hotspots, as both figures illustrate similar trends, and the increase in Fig. 4.4 leads ahead of similar increase in 4.2. We make two observations from Fig 4.4. Since only college counties experienced a high proportion of hotspots around mid-August, colleges which have much larger impact on college counties were the primary cause of the spike of hotspots. These findings are consistent with Lu et al.'s study.

On the other hand, it is not clear if college counties were consistently super spreaders, especially after 2-3 weeks of the Fall instruction [**25**] . However, if colleges were the major factor of the late high infection level the proportion of hotspots among non-college counties should have stayed relatively low compared to the proportion among college counties. Instead, the proportion among non-college counties was similar to the proportion of non-college counties after mid-September. The outbreaks in the fourth quarter were more likely to be driven by factors such as the seasonal trend and demographics instead of colleges.

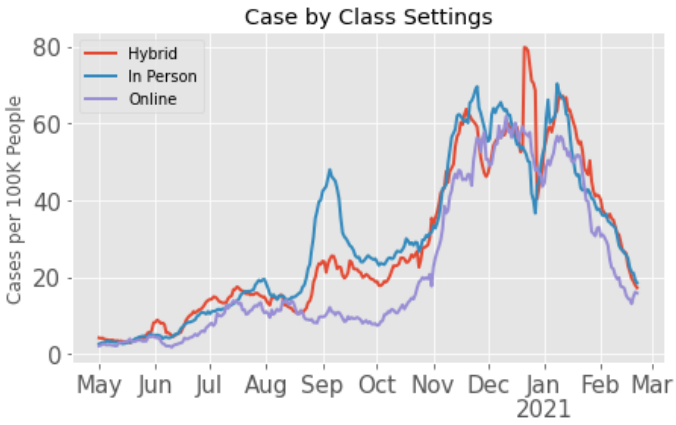


FIGURE 4.5. County cases per 100k people for counties with a single university in our study. We compare 7 day averages for counties, their matches, and the US total.

4.4. Association with College Testing Policies

We studied the association between college policies and hotspot activity of its home county. The results in this section are based on the association between the variables of interest and hotspot status and should not be taken as formal causal conclusions. To simplify the analysis and identify the effect of college policies, we only examine 92 college counties in which there is a single university (according to the College Scorecard dataset). We split counties into two categories which are hotspot and non-hotspot groups based on the label of each county on 14 August. We chose the date because the percentage of hotspot counties peaked on 14 August over the reopening period (August to September).

4.4.1. Effects of College Policies on Hotspots. The college response data that we collected contains testing-related information of 290 colleges. We obtained the information by manually accessing and combing through institution websites. In some instances, the colleges may have offered these resources but it was not displayed on their website. The information is summarized into five columns: class setting (e.g. hybrid), testing conducted (if the college conducted COVID-19 tests at all), testing availability (if the university provides COVID-19 testing), testing type (e.g. symptomatic) and test performed (e.g. nasal swab).

Figure 4.6 illustrates the distribution of each categorical variable. For each variable, We form a contingency table with the county group variable (hotspot and non-hotspot). Then, we conducted

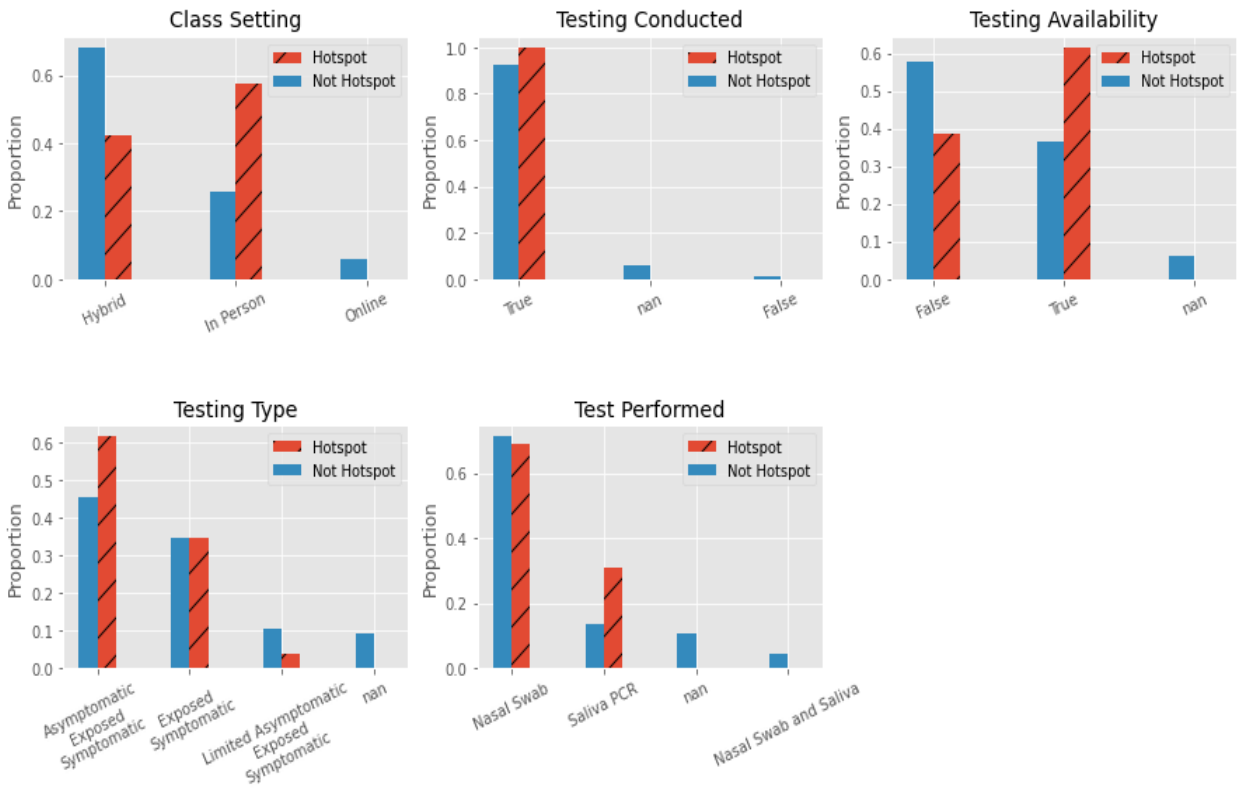


FIGURE 4.6. Bar Charts of College Policy Variables

the Fisher exact test of independence on each table. When we conducted the Fisher exact test for the Class Setting variable, we convert "In Person to Hybrid" and "In Person to Remote" to "In Person" for two reasons. First, these two categories only have three institutions in total. Next, we are interested in the policies around the start of Fall instruction. These three institutions initially were probably having in-person instruction. So we changed these levels to "In Person" rather than "Hybrid" or "Online".

	Class Setting	T Conducted	T Availability	T Type	T Performed
P-value	0.0157	0.6952	0.0634	0.2970	0.0861

TABLE 4.3. Summary of Fisher Exact Test of Independence. T stands for Testing or Test.

With the level of tests being 0.05 and Bonferroni correction, none of these tests yielded significant results as in Table 4.3 illustrates. However, the test of the Class Setting variable is on

the borderline of being significant and the bar chart shows that hotspot counties had a higher percentage of in-person instruction. Therefore, we further examined the relationship between cases and class settings. Figure 4.5 shows that around the time of college reopening in Fall, There are clear gaps between cases of counties with different class settings. So despite the non-significant result from the hypothesis testing, class setting might still impact the severity of the outbreaks at college counties at the beginning of Fall instruction. However, gaps between cases of counties with different class settings were less obvious since mid-October. As we argued in Section 4.3.2, there is little evidence that college students disproportionately contributed to outbreaks after the beginning of the Fall instruction.

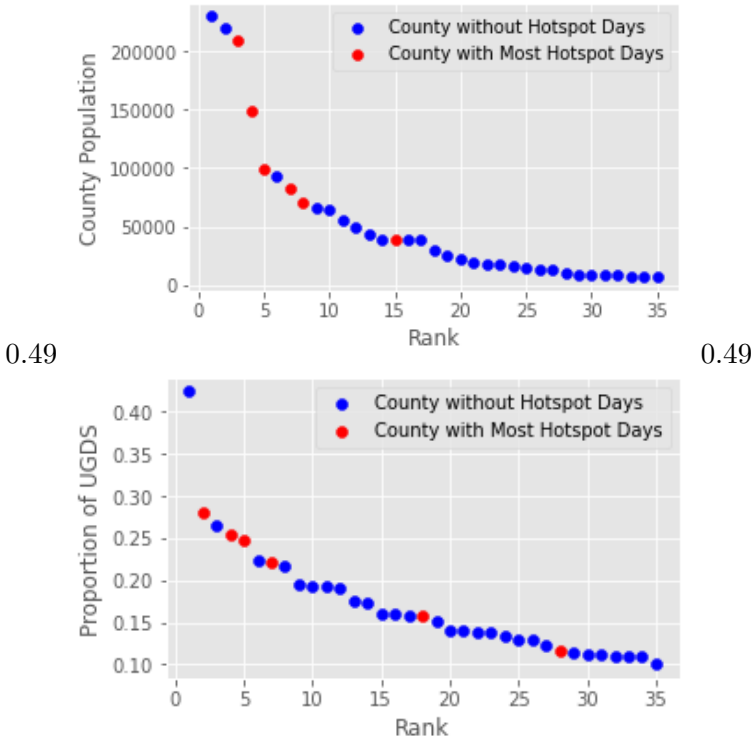


FIGURE 4.7. The college county populations (left) and proportion of undergraduate degree seeking students (right) in descending order. We highlight the 6 counties with the most hotspot days.

4.4.2. Case Studies. In this section, we will take a closer look at six college counties which suffered the worst outbreaks from August to September and college counties without hotspots over the same period. Specifically, we selected every four dates from 6 August to 15 September (11

days) and checked if a county was a hotspot on each day. A county with many days as a hotspot indicates that cases at the county were consistently increasing. Conversely, a county with few or no days as a hotspot underwent either decreasing or stable COVID cases. We rank the outbreak level at college counties by the number of hotspot days.

Table 4.4 shows that counties with most hotspot days don't have obvious geographical patterns. Similarly, Table 4.5 shows the counties with no hotspot days and there is no clear geographical pattern. The six worst performing college counties had testing available to the students (either nasal swab or saliva testing). In addition, Figure 4.7 shows that these counties have relatively high population and percentages of undergraduate students. The apparent association between the county population and hotspot number indicated that population is an important additional factor. However, the association between the proportion of undergrads and the hotspot number supports the conclusion that students returning to campus does contribute to hotspots.

There seem to be a number of factors contributing to the severity of outbreaks within college counties. We have reason to believe that the class setting contributes to COVID-19 hotspots, but it is one of several potential factors. We found no evidence that testing availability and the types of tests performed was associated with hotspots. However, we are dramatically limited by the amount of reliable data on hand, and with more data we may see stronger evidence of association.

State	County	University	Hotspot #	Class Setting
WY	Albany	U. of Wyoming	8	Hybrid
IL	Champaign	U. of Illinois Urbana-Champaign	7	In Person
IN	Monroe	Indiana U.-Bloomington	7	Hybrid
OK	Payne	Oklahoma State U.-Main	7	In Person
VA	Montgomery	Virginia Tech	7	Hybrid
WI	Portage	U. of Wisconsin-Stevens Point	7	Hybrid

TABLE 4.4. 6 Counties with Most Hotspot Days (out of 11 Days). All of them conducted testing and testing are available to students who want to get tested.

4.5. Discussion

We are able to draw several tentative conclusions from this analysis. First, there was a substantial increase in reported COVID-19 cases among college counties relative to similar counties. This increase happened during the college re-opening period, but after this initial period the cases

State	County	University	Class Setting
AL	Macon County	Tuskegee University	Hybrid
AL	Sumter County	University of West Alabama	In Person
AR	Pope County	Arkansas Tech University	In Person
CA	Yolo County	University of California-Davis	Online
CO	Alamosa County	Adams State University	In Person
CO	Gunnison County	Western Colorado University	Hybrid
GA	Emanuel County	East Georgia State College	In Person
GA	Lamar County	Gordon State College	Hybrid
IA	Decatur County	Graceland University-Lamoni	Hybrid
KY	Calloway County	Murray State University	Hybrid
KY	Campbell County	Northern Kentucky University	Hybrid
LA	Natchitoches Parish	Northwestern State University of Louisiana	Hybrid
MO	Adair County	Truman State University	Hybrid
MS	Claiborne County	Alcorn State University	Hybrid
MS	Jefferson County	Alcorn State University	Hybrid
MT	Beaverhead County	The University of Montana-Western	Hybrid
ND	Traill County	Mayville State University	Hybrid
NE	Nemaha County	Peru State College	Hybrid
OK	Cherokee County	Northeastern State University	Hybrid
PA	Clarion County	Clarion University of Pennsylvania	Online
SD	Lake County	Dakota State University	Hybrid
TX	Brewster County	Sul Ross State University	In Person
TX	Erath County	Tarleton State University	Hybrid
TX	Hays County	Texas State University	Hybrid
TX	Nacogdoches County	Stephen F Austin State University	Hybrid
UT	Iron County	Southern Utah University	Hybrid
VA	Buena Vista city	Southern Virginia University	Hybrid
VA	Fredericksburg city	University of Mary Washington	Hybrid
VA	Williamsburg city	William & Mary	Hybrid

TABLE 4.5. 29 College Counties with no Hotspot Days (out of 11 Days)

in college counties decreases and tracked closely with their matched counterparts. The increase was not due to the randomness of the case counts and we see that during the reopening the proportion of college counties that were hotspots is nearly 4 times the number in their matched counterparts. We also see that while there is little evidence to suggest that testing availability and test type is associated with hotspot activity, there is some evidence to suggest the class setting (in-person, remote, hybrid) is associated with case counts (but not significant at the 5% level with multiple testing correction).

The COVID-19 death rates do not significantly differ between college counties and their matched counterparts. Because the matched counties also had similar proportions of seniors, we conclude that there is little evidence to show that hotspots at the time of college reopening spread to the remainder of the community. If they had, it should have increased the death rate within the college counties as well, relative to their demographically matched counterparts.

This analysis was greatly constrained by the availability and granularity of COVID-19 data reporting. The county level is the finest spatial granularity at which we can obtain reliable and comparable case and death counts across states. Because the analysis was done at the county level, we were only able to identify 152 counties that were "college counties" (out of the 3,006 counties in the US). Of these counties only 92 could be matched to one primary university, which we can compare to the 1,625 4-year universities in the US. Imagine if reporting on COVID-19 cases was done at the census tract level—typically census tracts have between 1000 and 8000 people—then we might be able to pinpoint precisely which universities saw COVID-19 outbreaks. Furthermore, just like colleges that accept federal aid are required to report certain statistics to the US Department of Education (as in the College Scorecard dataset), such colleges could be mandated to report public health measures taken, such as in-person, remote, or hybrid classes in the case of an outbreak.

4.6. Appendix

4.6.1. College Counties. We didn't include Salt Lake County, UT in the analysis, though based on data we used, it satisfied our criterion of a college county. We exclude the Salt Lake County because the college with most undergraduate students is Western Governors University which only provides online education. Thus, most students may not reside in Salt Lake County. If we exclude the Western Governors University, Salt Lake County is no longer a college county by our definition.

4.6.2. Matching process. The matching problem can be formulated as an assignment problem was solved using the SciPy python package. Let d_i and a_i denote the standardized (zero mean and variance 1) population density and proportion of the senior population of county i. Then $x_i = (d_i, a_i)$ is the tuple of county i. Define an assignment function $f : C_s \to N_s$ where C_s and N_s are the set of indices of college and non-college counties of state s. The assignment problem seeks

to find the f which minimize

$$\sum_{i \in N_s} \|x_i - x_{f(i)}\|_2$$

for each state, where $\|.\|_2$ is Euclidean distance.

4.6.3. Hotspot definition. Let's introduce some notations which will be used in our definition of hotspots.

- C_i^t: count of COVID-19 cases of county i within 14 days after the pivot date t (exclude the pivot date).
- C_i^{t-1}: count of COVID-19 cases of county i within 14 days before the pivot date t (include the pivot date).
- P_i: population of county i.

Capital letters represent random variables and lowercase letters represent observed values. Our model assumes that $C_i^t \sim$Poisson(λ_i^t) and C_i^t's are independent for all t's and i's.

Throughout, we assume the population is constant over time. This assumption is unlikely to hold, especially for college counties, however an accurate estimate of the monthly population is difficult due to the infrequency of the US census and the paucity and biased-ness of realtime tracking information. In fact, it is well known that the problem of estimating seasonal population is especially difficulty with few satisfactory approaches [**16**].

We define county i to be a hotspot at time t if cases per capita increases 10% or more compared to time $t-1$. Formally, we define the following hypothesis testing problem for the i, t pair,

$$H_0^{i,t} : \frac{\frac{\lambda_i^t}{P_i}}{\frac{\lambda_i^{t-1}}{P_i}} \leq 1.1 \; vs. \; H_1^{i,t} : \frac{\frac{\lambda_i^t}{P_i}}{\frac{\lambda_i^{t-1}}{P_i}} > 1.1.$$

The rejection of H_0 will identify county i to be a hotspot at time t. Since we assume the population of each county doesn't vary over time, the hypothesis can be simplified to the following:

$$H_0^{i,t} : \frac{\lambda_i^t}{\lambda_i^{t-1}} \leq 1.1 \; vs. \; H_1^{i,t} : \frac{\lambda_i^t}{\lambda_i^{t-1}} > 1.1. \tag{4.1}$$

We will address multiple hypothesis corrections in the following section, but we first need a p-value for each individual test. We use the conditional test (C-Test) as mentioned in Section 2 of [**21**] to compute the p-value. The idea behind the C-test is the following. Let $p_i^t = \frac{\lambda_i^t}{\lambda_i^t + \lambda_i^{t-1}}$ and $N_i^t = C_i^t + C_i^{t-1}$. Since $C_i^t | N_i^t \sim$Binomial(N_i^t, p_i^t) under H_0, so we may compute the p-value for the monotone likelihood ratio test using the CDF of the binomial distribution with success probability of $\frac{1}{1+1.1} \approx 0.476$ as the null hypothesis.

In summary, rejecting the null hypothesis in (4.1) means that there is a substantial increase in case incidences at time t. Notably, it does not mean that the county has a high case count, and we may have a high case count without rejecting the null hypothesis if if it consistently had high cases per capita during time period $t-1$ and t. It remains to translate this problem as a multiple hypothesis test.

4.6.4. Multiple Hypothesis Testing and FDR Correction. We carry out the C-test for all 3143 counties, date pairs at an equispaced (4 days) sequence of dates between 1st May, 2020, and 21st Feb, 2021. If we do not take multiple testing into account, we are bound to reject a large proportion of null hypotheses. Therefore, we adopt the Benjamini–Hochberg (BH) procedure [**2**] to control the false discovery rate (FDR) at level $\alpha = 0.05$. The FDR is defined to be the percentage of the true null hypotheses that are falsely rejected.

Let us provide a recap of the BH procedure. The procedure assumes that all p-values are independent. Then

(1) Let $P_{(1)} \leq P_{(2)}, \ldots, P_{(m)}$ be the ordered p-values.
(2) Let $I_i = \frac{i}{m}\alpha$ and $T = \max\{i : I_i < P_{(i)}\}$.
(3) Reject all H_0's where $P_{(i)} \leq P_{(T)}$.

The BH procedure is guaranteed to control FDR at level α assuming that the tests are independent, although there are weaker conditions in which FDR is controlled [**3**]. Our hope is that by controlling FDR using the BH procedure, the hypothesis testing can be less conservative compared to using the Bonferroni correction.

Bibliography

[1] Abbott, S., Hellewell, J., Thompson, R. N., Sherratt, K., Gibbs, H. P., Bosse, N. I., Munday, J. D., Meakin, S., Doughty, E. L., Chun, J. Y., et al. Estimating the time-varying reproduction number of sars-cov-2 using national and subnational case counts. *Wellcome Open Research 5*, 112 (2020), 112.

[2] Benjamini, Y., and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57*, 1 (1995), 289–300.

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

[4] Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*, 1 (2020), 175–197.

[5] Bureau, U. C. County population totals: 2010-2019. `https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html`, 2019.

[6] Candes, E., Fan, Y., Janson, L., and Lv, J. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection, 2016.

[7] Casella, G. An introduction to empirical bayes data analysis. *The American Statistician 39*, 2 (1985), 83–87.

[8] Csiszar, I., and Matus, F. Information projections revisited. *IEEE Transactions on Information Theory 49*, 6 (2003), 1474–1490.

[9] Dawid, A. P. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological) 41*, 1 (1979), 1–31.

[10] Dong, E., Du, H., and Gardner, L. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases 20*, 5 (2020), 533–534.

[11] Drton, M., and Maathuis, M. H. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application 4*, 1 (2017), 365–393.

[12] Dua, D., and Graff, C. UCI machine learning repository, 2017.

[13] Farasat, A., Nikolaev, A., Srihari, S. N., and Blair, R. H. Probabilistic graphical models in modern social network analysis. *Social Network Analysis and Mining 5* (10 2015).

[14] Farrow, D. C., Brooks, L. C., Rumack, A., Tibshirani, R. J., and Rosenfeld, R. Delphi Epidata API. `https://github.com/cmu-delphi/delphi-epidata`, 2015.

[15] Gao, C., Zheng, Y., Wang, W., Feng, F., He, X., and Li, Y. Causal inference in recommender systems: A survey and future directions, 2022.

[16] Happel, S. K., and Hogan, T. D. Counting snowbirds: The importance of and the problems with estimating seasonal populations.

[17] Huang, T.-M. Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics 38*, 4 (2010), 2047–2091.

[18] Ivory, D., Gebeloff, R., and Mervosh, S. Young people have less covid-19 risk, but in college towns, deaths rose fast. *The New York Times*.

[19] Katsevich, E., and Ramdas, A. On the power of conditional independence testing under model-x, 2020.

[20] Kline, P., and Santos, A. A score based approach to wild bootstrap inference. *Journal of Econometric Methods 1*, 1 (2012), 23–41.

[21] Krishnamoorthy, K., and Thomson, J. A more powerful test for comparing two poisson means. *Journal of Statistical Planning and Inference 119*, 1 (2004), 23–35.

[22] Li, C., and Fan, X. On nonparametric conditional independence tests for continuous variables. *WIREs Computational Statistics 12*, 3 (2020), e1489.

[23] Liu, R. Y. Bootstrap Procedures under some Non-I.I.D. Models. *The Annals of Statistics 16*, 4 (1988), 1696 – 1708.

[24] Lu, A., Lu, A., Schormann, W., Ghassemi, M., Andrews, D., and Moses, A. The cells out of sample (coos) dataset and benchmarks for measuring out-of-sample generalization of image classifiers. In *Advances in Neural Information Processing Systems* (2019), H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc.

[25] Lu, H., Weintz, C., Pace, J., Indana, D., Linka, K., and Kuhl, E. Are college campuses superspreaders? a data-driven modeling study. *medRxiv* (2020).

[26] Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. A unifying view on dataset shift in classification. *Pattern Recognition 45*, 1 (2012), 521–530.

[27] Neykov, M., Balakrishnan, S., and Wasserman, L. Minimax optimal conditional independence testing, 2020.

[28] of Education, U. D. College scorecard. `https://collegescorecard.ed.gov/data/documentation/`, 2020.

[29] Ostrovskii, D., and Bach, F. Finite-sample analysis of m-estimators using self-concordance, 2018.

[30] Puli, A., Zhang, L. H., Oermann, E. K., and Ranganath, R. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations, 2021.

[31] Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences 63*, 10 (sep 2020), 1872–1897.

[32] Riddell, A., Hartikainen, A., and Carter, M. pystan (3.0.0). PyPI, Mar. 2021.

[33] Rosenbaum, P. R. Optimal matching for observational studies. *Journal of the American Statistical Association 84*, 408 (1989), 1024–1032.

[34] Rudelson, M., and Vershynin, R. Hanson-wright inequality and sub-gaussian concentration, 2013.

[35] Runge, J. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information, 2017.

[36] Safford, H., Zuniga-Montanez, R. E., Kim, M., Wu, X., Wei, L., Sharpnack, J., Shapiro, K., and Bischel, H. N. Wastewater-based epidemiology for covid-19: Handling qpcr nondetects and comparing spatially granular wastewater and clinical data trends. *ACS ESamp;T Water* (2022).

[37] Sen, R., Suresh, A. T., Shanmugam, K., Dimakis, A. G., and Shakkottai, S. Model-powered conditional independence test, 2017.

[38] Shah, R. D., and Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics 48*, 3 (jun 2020).

[39] Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

[40] Strobl, E. V., Zhang, K., and Visweswaran, S. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference 7*, 1 (2019), 20180017.

[41] Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. No. 47 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[42] Vershynin, R. Introduction to the non-asymptotic analysis of random matrices.

[43] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods 17* (2020), 261–272.

[44] Wakefield, J. *Bayesian and Frequentist Regression Methods*. Springer Series in Statistics. Springer New York, 2013.

[45] Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. Conditional distance correlation. *Journal of the American Statistical Association 110*, 512 (2015), 1726–1734.

[46] Wooldridge, J. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2002.

[47] Wu, C. F. J. Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics 14*, 4 (1986), 1261 – 1295.

[48] Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research 16*, 115 (2015), 3813–3847.

[49] Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. AUAI Press, pp. 804–813.

[50] Zhong, W., Dong, L., Poston, T. B., Darville, T., Spracklen, C. N., Wu, D., Mohlke, K. L., Li, Y., Li, Q., and Zheng, X. Inferring regulatory networks from mixed observational data using directed acyclic graphs. *Frontiers in Genetics 11* (2020).