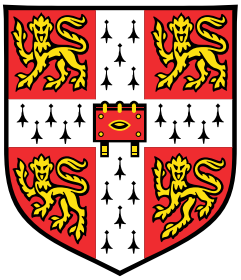
Hypothesis testing and causal inference with heterogeneous medical data



Alexis Bellot

Department of Applied Mathematics and Theoretical Physics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Darwin College

December 2020

While uncovering cause and effect is in general challenging, in this case it is obvious.

To my parents.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation does not exceed the prescribed word limit.

Alexis Bellot
December 2020

Acknowledgements

A great many persons have helped with the inspiring and sometimes difficult task of writing a dissertation, have shown unfailing support over the years, and have furthered my progress by stimulating suggestions and criticism. To all of them I offer heartfelt thanks. Here I shall mention by name only my advisor, Mihaela van der Schaar, who introduced me to the field of machine learning and its applications in medicine, made me believe in her vision for improving healthcare and supported me to explore new and varied research directions; my closest collaborators Yao Zhang, Trent Kyono, Zhaozhi Qian, Ahmed Alaa, Thomas Cowling and Yue Ruan, who were wonderful research partners and from whom I learned a great amount; and my family and friends, whose unconditional support I value tremendously. In many ways, as individuals we are simply the sum of our interactions and so is this dissertation, which would not have been possible without them.

Abstract

Learning from data which associations hold and are likely to hold in the future is a fundamental part of scientific discovery. With increasingly heterogeneous data collection practices, exemplified by passively collected electronic health records or high-dimensional genetic data with only few observed samples, biases and spurious correlations are prevalent. These are called spurious because they do not contribute to the effect being studied. In this context, the modelling assumptions of existing statistical tests and causal inference methods are often found inadequate and their practical utility diminished even though these models are increasingly used as decision-support tools in practice. This thesis investigates how modern computational techniques may broaden the fields of hypothesis testing and causal inference to handle the subtleties of large heterogeneous data sets, as well as simultaneously improve the robustness and theoretical understanding of machine learning algorithms using insights from causality and statistics.

The first part of this thesis is concerned with hypothesis testing. We develop a framework for hypothesis testing on set-valued data, a representation that faithfully describes many real-world phenomena including patient biomarker trajectories in the hospital. Using similar techniques, we develop next a two-sample test for making inference on selection-biased data, in the sense that not all individuals are equally likely to be included in the study, a fact that biases tests if not accounted for and if the desideratum is to obtain conclusions that are generally applicable. We conclude this section with an investigation of conditional independence in high-dimensional data, such as found in gene expression data, and propose a test using generative adversarial networks. The second part of this thesis is concerned with causal inference and discovery, with a special focus on the influence of unobserved confounders that distort the observed associations between variables and yet may not be ruled out or adjusted for using data alone. We start by demonstrating that unobserved confounders may bias substantially the generalization performance of machine learning algorithms trained with conventional learning paradigms such as empirical risk minimization. Acknowledging this spurious effect, we develop a new learning principle inspired by causal insights that provably generalizes to

test data sampled from a larger set of distributions different from the training distribution. In the last chapter we consider the influence of unobserved confounders for causal discovery. We show that with some assumptions on the type and influence on the nature of unobserved confounding one may develop provably consistent causal discovery algorithms, formulated as a solution to a continuous optimization program.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Symmetries are central to our understanding of nature. They define a certain invariance and regularity in the behaviour of a system when observed from different viewpoints. For instance, the ability to repeat experiments at different places and at different times is based on the invariance of the laws of nature under space-time translations. Symmetry principles provide structure and coherence to the laws of nature and by implication structure and coherence to experimental observations.

In every day life, it may seem obvious that systems or objects should not change when we change the perspective with which we observe them – and indeed we would be living in a very different world if this weren't the case – but the consequences with respect to the underlying laws of nature reach much further, and in fact the concrete mathematical implications of symmetries for the physical world were not shown until 1918. That year, Emmy Noether [123] demonstrated that for each symmetry that exists in nature there must also exist a corresponding quantity (such as the energy of a system or its momentum) that is conserved in time. In a formal sense, symmetry principles *dictate* the form of the laws of nature, a fact that quickly came to shape the discovery of new physical laws during the twentieth century. The attempt to use symmetries in nature to infer the fundamental laws of physics proliferated, and with it the hope that conserved quantities measured in experiments would allow one to work backward to find out the underlying laws of nature. The concept has become so powerful that in the words of Nobel laureate P. W. Anderson *"It is only slightly overstating the case to say that physics is the study of symmetry."* [2].

A similar excitement permeates the fields of Statistics and Machine Learning today. The fundamental relationships between different variables, when measured through independent

observations in well-designed experiments, leave a statistical footprint that characterizes the underlying data generating system. For instance, the fact that lung cancer is more often observed together with a history of smoking than otherwise suggests that there may be a more fundamental dependency between the two. Statistical covariance or dependence between measured variables may be used, just as symmetries in the natural world, to ground more fundamental relationships between variables and the events they represent. The data gathered may not only describe what happened in a finite number of experiments but transcend the actual observations made and point to a general property of an underlying data generating process, a property of the fundamental laws governing the system. Quantifying the evidence for a particular dependency structure in the distribution of data, and separating it from chance occurrences or coincidences, is the essence of hypothesis testing.

Hypothesis testing is the practice of defining a function of the data that provably discriminates between two statistical hypotheses of interest, such as the hypothesis that lung cancer and smoking are statistically independent variables, or the hypothesis that groups of smokers and non-smokers have the same distribution of lung cancer outcomes, for example. When hypotheses are well-posed and the data well-behaved, for instance having observations sampled from a known distributional family or collected independently from a well-designed experiment, hypothesis tests have demonstrated beyond reasonable doubt countless dependencies between variables that inform our everyday decisions, such as refraining from smoking to avoid lung lancer [26, 135].

However, even if the analogy between symmetries in physics and dependencies in data is compelling, our present understanding of data and its underlying data generating mechanism is rather incomplete. Beyond the difficulties of data recording and its inherent biases, not all questions regarding the structure of the data generating process can necessarily be answered with an estimated data distribution. Statisticians established an association between smoking and lung cancer in the 1950s but were not able, from data alone, to say with certainty whether forcing people to quit smoking subsequently reduces their risk for cancer. Questions of this type refer to likely distributions of outcomes *after* interventions (e.g. the distribution of lung cancer after forcing people to stop smoking), a system that may exhibit different dependencies between variables that observed without intervening. This type of inference is fundamentally not accessible without further assumptions on the causal relationships between variables.

Causal associations, as opposed to statistical association, describe a kind of relationship between variables in a system that supports reasoning about the consequences of variable manipulations, and about counterfactual scenarios. It is one that describes a system under changing environments and one that supports human explanation as it mirrors the way humans

model the world (through cause and effect). A causal model can be thought of as a collection of interventional or counterfactual distributions in contrast to statistical models that may be defined as a collection of observational distributions. Data has much to contribute to the study of causality. For instance, precise hypotheses may be clearly formulated in causal diagrams whose consequences may be tested and compared to statistical constraints in observational data [129]. For a given causal graph, one may build algorithms to quantify causal relationships from data – adjusting for explicit confounding factors – and one may even discover the causal structure from scratch with explicit assumptions on the relationship between statistical and causal associations. The study of dependencies in data and its connections to causality is increasingly important as data science is applied in decision-making contexts.

Just as symmetries revolutionized our understanding of physics, causal inference and hypothesis testing conducted on increasingly large and heterogeneous datasets, from medicine, economics, sociology or climate science, promise to uncover deeper insights about the nature of variable interactions and underlying processes. An important challenge, however, is that modern datasets do not lend themselves naturally to the requirements of current hypothesis testing and causal inference methods. Modern datasets are passively collected, heterogeneous, and biased in a myriad of ways with measurement errors, missingness and inconsistencies. Unless properly accounted for, potential biases may call any data-driven conclusion into question even in the infinite data regime.

1.1 Contributions

This dissertation advances the state of the art of hypothesis testing and causal inference. It highlights some of these biases and develops tools and algorithms that ensure correct conclusions in the presence of biased data. The motivation for the proposed techniques lie in specific problems in medicine and biology, where these problems are acute, yet the potential for improving care in a data-driven, individualized fashion is enormous.

In the following paragraphs I summarize the contributions presented in each of the subsequent chapters, and list all author publications in Table 1.1. All chapters are self-contained, divided into investigations of hypothesis testing in the first part of this dissertation and into investigations of the causal structure of data and its benefits for robust prediction and inference in a second part. Some overlap in the background covered in each chapter is present, but allows for reading each chapter independently of each other and should not disturb the overall flow of the dissertation. We invite the reader to explore each chapter in no particular order.

1.1.1 Kernel hypothesis testing for set-valued data

In the second chapter, we present a general framework for kernel hypothesis testing on distributions of *sets* of individual examples. Sets may represent many common data sources such as groups of observations in time series, collections of words in text or a batch of images of a given phenomenon. This observation pattern, however, differs from the common assumptions required for hypothesis testing: each set differs in size, may have differing levels of noise, and also may incorporate nuisance variability, irrelevant for the analysis of the phenomenon of interest; all features that bias test decisions if not accounted for. We propose to interpret sets as independent samples from a collection of latent probability distributions, and introduce kernel two-sample and independence tests in this latent space of distributions. We prove the consistency of these tests and observe them to outperform in a wide range of synthetic and real data experiments, where previously heuristics were needed for feature extraction and testing.

1.1.2 A kernel two-sample test with sample selection bias

In the third chapter, we propose new test that acknowledges for bias in the data collection mechanism. Hypothesis tests, like other data driven methods, may inherit biases embedded in the data collection mechanism (some instances often being systematically more likely included in our sample) and consistently reproduce biased decisions. Our contribution is a two-sample test that adjusts for selection bias by accounting for differences in marginal distributions of confounding variables. Our test statistic is a weighted distance between samples embedded in a reproducing kernel Hilbert space, whose balancing weights provably correct for certain kinds of bias. As in other chapters, we conclude with controlled experiments that highlight the benefit of this adjustment and explore the use of our test on treatment effect studies from economics.

1.1.3 Conditional independence testing using adversarial neural networks

In the fourth chapter, we consider the hypothesis testing problem of detecting conditional dependence, with a focus on high-dimensional feature spaces, such as may be encountered in gene expression data. Our contribution is a new test statistic based on samples from a generative adversarial network designed to approximate directly a conditional distribution that encodes the null hypothesis, in a manner that maximizes power (the rate of true negatives). We show that such an approach requires only that density approximation be viable in order to ensure that we control type I error (the rate of false positives); in particular, no assumptions need to

be made on the form of the distributions or feature dependencies. Using synthetic simulations with high-dimensional data we demonstrate significant gains in power over competing methods. In addition, we illustrate the use of our test to discover causal markers of disease in genetic data.

1.1.4 Accounting for unobserved confounding in domain generalization

The fifth chapter starts the second part of this dissertation investigating the influence of bias due to *unobserved* confounders on the prediction generalization performance of common learning principles such as empirical risk minimization. We argue for defining generalization with respect to a broader class of distribution shifts (defined as arising from interventions in the underlying causal model), including changes in observed, unobserved and target variable distributions. Our contribution is a new robust learning principle that may be paired with any gradient-based learning algorithm. This learning principle has explicit generalization guarantees, and relates robustness with certain invariances in the causal model, clarifying why, in some cases, test performance lags training performance.

1.1.5 Scoring DAGs with dense unobserved confounding

Unobserved confounding is also one of the greatest challenges for causal discovery. The case in which unobserved variables have a potentially widespread effect on many of the observed ones is particularly difficult because most pairs of variables are conditionally dependent given any other subset. In the sixth chapter we show that beyond conditional independencies, unobserved confounding in this setting leaves a characteristic footprint in the observed data distribution that allows for disentangling spurious and causal effects. Using this insight, we demonstrate that a sparse linear Gaussian directed acyclic graph among observed variables may be recovered approximately and propose an adjusted score-based causal discovery algorithm that may be implemented with general purpose solvers and scales to high-dimensional problems. We find, in addition, that despite the conditions we pose to guarantee causal recovery, performance in practice is robust to large deviations in model assumptions.

Table 1.1: Author publications and preprints.

Scoring DAGs with Dense Unobserved Confounding **A Bellot**, M van der Schaar, 2020
Accounting for Unobserved Confounding in Domain Generalization **A Bellot**, M van der Schaar, 2020
Kernel Hypothesis Testing with Set-valued Data **A Bellot**, M van der Schaar. *Conference on Uncertainty in Artificial Intelligence*, 2021
AI-based Hypothesis Testing in Individuals with CF **A Bellot**, R A Floto, M van der Schaar, Pediatric Pulmonology (Abstract) 2020
Logistic regression and machine learning predicted patient mortality from large sets of diagnosis codes comparably Cowling, T. E., **Bellot, A**. , and others. *Journal of Clinical Epidemiology*, 2020
A Kernel Two-Sample Test for Unbiased Decisions **A Bellot**, M van der Schaar, *Conference on Uncertainty in Artificial Intelligence*, 2021
One-year mortality of colorectal cancer patients: development and validation of a prediction model Cowling, T. E., **Bellot, A**., Boyle, J., and others. *British Journal of Cancer*, 2020
Predicting the Risk of Inpatient Hypoglycemia With Machine Learning using Electronic Health Records Y Ruan, **A Bellot**, Z Moysova, GD Tan, A Lumb, and others. *Diabetes care*, 2020
Flexible Modelling of Longitudinal Medical Data: A Bayesian Nonparametric Approach **A Bellot**, M van der Schaar, *ACM Transactions on Computing for Healthcare*, 2020
Learning overlapping representations for the estimation of individualized treatment effects Y Zhang, **A Bellot**, M van der Schaar, *AISTATS*, 2020
Learning Dynamic and Personalized Comorbidity Networks from Event Data using Deep Diffusion Processes Z Qian, AM Alaa, **A Bellot**, J Rashbass, M van der Schaar, *AISTATS*, 2020
Conditional Independence Testing using Generative Adversarial Networks **A Bellot**, M van der Schaar *Advances in Neural Information Processing Systems*, 2019
Boosting transfer learning with survival data from heterogeneous domains **A Bellot**, M van der Schaar, *AISTATS*, 2019
Multitask boosting for survival analysis with competing risks **A Bellot**, M van der Schaar, *Advances in Neural Information Processing Systems*, 2018
Boosted trees for risk prognosis **A Bellot**, M van der Schaar, *Machine Learning for Healthcare Conference*, 2018
A hierarchical bayesian model for personalized survival predictions **A Bellot**, M Van der Schaar, *IEEE journal of biomedical and health informatics*, 2018
Tree-based bayesian mixture model for competing risks **A Bellot**, M van der Schaar, *AISTATS*, 2018

Part I

Hypothesis Testing with Heterogeneous Data

Chapter 2

Kernel Hypothesis Testing with Set-valued Data

In this chapter, we present a general framework for kernel hypothesis testing on distributions of sets of individual examples. Sets may represent many common data sources such as groups of observations in time series, collections of words in text or a batch of images of a given phenomenon. This observation pattern, however, differs from the common assumptions required for hypothesis testing: each set differs in size, may have differing levels of noise, and also may incorporate nuisance variability, irrelevant for the analysis of the phenomenon of interest; all features that bias test decisions if not accounted for. In this chapter, we propose to interpret sets as independent samples from a collection of latent probability distributions, and introduce kernel two-sample and independence tests in this latent space of distributions. We prove the consistency of these tests and observe them to outperform in a wide range of synthetic and real data experiments, where previously heuristics were needed for feature extraction and testing.

Introduction

Hypothesis tests are used to answer questions about a specific dependency structure in data (e.g. independence between variables, equality of distributions between samples etc.). They are used in applications across the sciences where they serve as an essential tool to summarize experimental data and quantify the evidence for discoveries on the relationship of variables of interest [103]. As a consequence, a growing body of work is constantly revisiting established modelling assumptions to allow for consistent testing in increasingly heterogeneous data sources. Examples include non-parametric tests formulated as distances in Hilbert space [73, 72, 70, 198], tests based on neural network representations [108, 114, 11] and others that

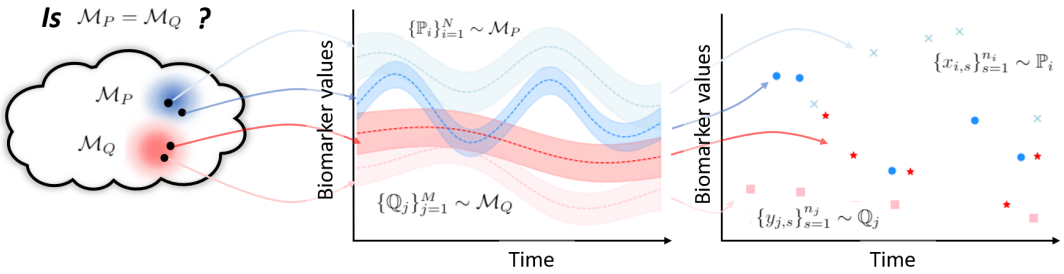


Figure 2.1: We consider an example from electronic health records to illustrate the proposed approach. **Right panel:** we observe irregular, uncertain biomarker measurements over time in two groups of patients (treated and control) colored with different shades of red and blue, the question being whether these populations have the same trajectory in distribution. **Middle panel**: we encode the uncertainty in each patient trajectory by a probability distributions on the space of observations. **Left panel:** The two-sample problem is to test for equality in distribution on the space of patient-specific distributions, rather than actual observations. This two-level hierarchy allows for noisy inputs and irregular input sizes. A description of the notation and more details can be found in Section 2.2.1.

have significantly advanced the reach of hypothesis tests towards high-dimensional data of unknown distribution.

Almost universally however, non-parametric tests require a *fixed* presentation of data (e.g. each instance living in $\mathbb{R}^d$) and do not account for *non-homogeneous noise* patterns across examples. Many problems do exhibit these properties, for example with medical data, where each patient has different levels of variation and have observations irregularly measured over time. A similar pattern is observed in many other domains involving time series and bagged data (e.g. multiple images of the same phenomenon).

Intriguingly, there exists an appropriate representation of data that naturally encodes a more flexible observation pattern, namely each example represented as a *set* of observations (i.e. an unordered collection of multivariate observations), each set of potentially irregular length and sampled from potentially different distributions. In particular, sets do not presuppose a fixed representation of data (sets may be of different length) and each set may be associated with a unique distribution that encodes its particular variation pattern (potentially different from other sets). Testing on sets implicitly shifts the question of interest from a hypothesis on groups of actual observations to an hypothesis on groups of latent distributions assumed to represent each observed example or set. See Figure 2.1 for an illustration of this interpretation for the two sample problem. This set-up is common in regression problems where one seeks to learn a mapping from distributions to associated labels [170, 171], but is unexplored in hypothesis

testing. The goal of this chapter is to introduce kernel two-sample and kernel independence tests defined on *set*-valued examples.

We will show that tests defined in this space appropriately encode individual-level heterogeneity, are much more flexible, do not require heuristic pre-processing of data, and are found to be more powerful than alternatives. We propose an approach applicable to any kernel-based test that includes, in addition to two-sample and independence tests described here, conditional independence tests and three-variable interaction tests.

The technical challenge to achieve consistency of test decisions is that latent distributions on which tests are defined are not available (and instead are approximated with each available set of observation). This introduces an additional layer of uncertainty that must be bounded to derive well-defined asymptotic distributions for the proposed test statistics. For this reason, we put emphasis also on the quality of finite-dimensional approximations of the proposed tests, with approaches to minimize test statistic variance and to tune hyperparameters for maximum power.

Our contributions are three-fold:

1. We formally describe tests on set-valued data, and to the best of our knowledge for the first time.
2. We demonstrate the consistency of these tests for the two-sample and independence testing problems.
3. We validate the proposed tests and optimization routines on simulated experiments that show that one may consistently discriminate between hypotheses on data that was previously not amenable to hypothesis testing.

2.1 Background

The tests presented in this chapter are formally defined on distributions. Testing on distributions is the problem of defining a test statistic that maps distributions to a scalar that quantifies the evidence for a hypothesis we might set on the relationships in data. However, we do not have access to probability distributions themselves, but rather distributions are observed only through *sets* of samples,

$$\{x_{1,j}\}_{j=1}^{n_1}, \ldots, \{x_{N,j}\}_{j=1}^{n_N}. \tag{2.1}$$

Each $\{x_{i,j}\}_{j=1}^{n_i}$ is a *set* of n_i individual observations $x_{i,j}$ (typically in $\mathbb{R}^d$). We assume that $\{x_{i,j}\}_{j=1}^{n_i}$ are *i.i.d* samples from an unobserved probability distribution $\mathbb{P}_i$. The probability distributions $\{\mathbb{P}_i\}_{i=1}^{N}$ themselves have inherent variability, such as can be expected for example from different medical patients. We assume each one of them to be drawn randomly from some unknown meta-distribution $\mathcal{M}_P$ defined over a set of probability measures $\mathcal{P}$. We illustrate this set-up in Figure 2.1 for the two-sample problem (more details in Section 2.2.1).

2.1.1 Embeddings of Distributions

Let $\mathcal{X}$ be a measurable space of observations. We use a positive definite bounded and measurable kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to represent distributions $\mathbb{P}_i$ on $\mathcal{X}$, and independent samples $\{x_{i,j}\}_{j=1}^{n_i}$, as two functions $\mu_{\mathbb{P}_i}$, and $\hat{\mu}_{\mathbb{P}_i}$ respectively, called kernel mean embeddings [121]. Both are defined in the corresponding Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ by,

$$\mu_{\mathbb{P}_i} := \int_{\mathcal{X}} k(x,\cdot) d\mathbb{P}_i(x), \qquad \hat{\mu}_{\mathbb{P}_i} := \frac{1}{n_i} \sum_{x \in \{x_{i,j}\}_{j=1}^{n_i}} k(x,\cdot).$$

To make inference on populations of distributions, the desiratum however is on defining useful representations of distributions $\mathcal{M}_P$ on the space probability measures, rather than on the space of observations. [38] showed that one may do so analogously to the definition of kernels on $\mathcal{X}$ by treating mean embeddings $\mu_{\mathbb{P}}$ themselves as inputs to kernel functions (replacing $x \in \mathcal{X}$ in the conventional learning setting as inputs to k). See eq. (2.2) below.

Accounting for variance in embedding approximations. In practice, each set representation $\mu_{\mathbb{P}_i}$ is limited to be approximated by irregularly sampled observations $\{x_{i,j}\}_{j=1}^{n_i}$. Not all mean embeddings $\mu_{\mathbb{P}}$ are expected to provide the same amount of information about their underlying distribution $\mathbb{P}$. Indeed, the empirical mean embeddings $\hat{\mu}_{\mathbb{P}_i}$ converge to their population counterpart at a rate $\mathcal{O}(1/\sqrt{n_i})$ (see e.g. Lemma 1 in the Appendix of this chapter and also [164]) in their set size n_i. Rather than assuming access to a uniform sample of distributions $\{\mathbb{P}_i\}_{i=1}^{N}$ from $\mathcal{M}_P$, like we did with the raw observations $\{x_{i,j}\}_{j=1}^{n_i}$, we may account for this irregularity and uncertainty in approximation by interpreting the set of distributions as a weighted sample $\{(\mathbb{P}_i, w_i)\}_{i=1}^{N} \sim \mathcal{M}_P$. Each weight quantifying the accuracy of the approximation of each distribution with the limited samples available. The corresponding population and empirical mean embedding in this space may be written as,

$$\mu_{\mathcal{M}} := \int_{\mathcal{P}} K(\mu_{\mathbb{P}},\cdot) d\mathcal{M}(\mathbb{P}), \qquad \hat{\mu}_{\mathcal{M}} := \sum_{i=1}^{N} w_i K(\mu_{\mathbb{P}_i},\cdot). \tag{2.2}$$

We will make use of the Gaussian kernel between distributions defined $K(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}}) := \exp(-||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||^2_{\mathcal{H}_K}/2\sigma^2)$ [38, 120]. Note that for kernels on $\mathcal{X}$, their RKHS consists of functions $\mathcal{X} \to \mathbb{R}$, while the kernel K lives on the space of distributions on $\mathcal{X}$, $\mathcal{P}(\mathcal{X})$, and its RKHS consists of functions $\mathcal{P}(\mathcal{X}) \to \mathbb{R}$. We may use K to learn from samples that are individual distributions, rather than individual observations [38].

Relationships with learning on distributions. With this construction (i.e. kernels evaluated on mean embeddings) [170] investigated generalization performance in distributional regression: regressing to a real-valued response from a probability distribution. Results that were subsequently extended to study distributional regression for causal inference [112] and for transfer learning [24]. A technical contribution of this chapter is to extend these results to demonstrate consistent hypothesis testing on distributions.

2.1.2 Hypothesis Testing with Kernels

The advantage for hypothesis testing of mapping distributions $\mathcal{M}$ and $\mathcal{M}'$ to functions in an RKHS is that we may now say that $\mathcal{M}$ and $\mathcal{M}'$ are close if the RKHS distance $||\mu_{\mathcal{M}} - \mu_{\mathcal{M}'}||_{\mathcal{H}_K}$ is small [70]. This distance depends on the choice of the kernel K and k; a crucial property of the embeddings is that for certain kernels the feature map is injective. These kernels are called characteristic [165]. Probability distributions may be distinguished exactly by their images in the RKHS, and also $||\mu_{\mathcal{M}} - \mu_{\mathcal{M}'}||_{\mathcal{H}_K}$ is zero if and only if the distributions coincide [70]. From the statistical testing point of view, this coincidence axiom is key as it ensures consistency of comparisons for any pair of different distributions.

As a key property of the set-up we have introduced, in Theorem 2.2 [38] demonstrated that for well known kernels, such as the Gaussian kernels, if used in both levels of the embedding and defined on a compact metric space the resulting embedding is injective (i.e. kernels are characteristic)[1].

The empirical version of the RKHS distance, however, will not necessarily be exactly zero even if the distributions do coincide. Some variability is to be expected due to the limited number of samples, and in contrast to conventional kernel tests, in the case considered here also due to the variability in the estimation of set embeddings. Instead of testing on an $i.i.d.$ sample $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$, we are testing over the set $\{\hat{\mu}_{\mathbb{P}_i}\}_{i=1}^N$. There is an additional level of uncertainty which must be accounted for.

[1] Theorem 2.2 [38] technically shows that such kernels are universal, but universal kernels on compact metric space are known to be characteristic, as shown in Theorem 1 [70].

In practice, tests are constructed such that a certain hypothesis is rejected whenever a test statistic exceeds a certain threshold away from 0 [103]. Then, short from achieving perfect discrimination between two hypotheses, the goal of hypothesis testing is to derive a threshold such that false positives are upper bounded by a design parameter α and false negatives are as low as possible.

2.1.3 Related work

Two-sample and independence testing are two of the most commonly used of all statistical procedures. Classical approaches to these problems apply to univariate data with high power only on restricted classes of alternative hypotheses. These include for the two-sample problem: Hotelling's two-sample t-squared statistic [79], Kolmogorov-Smirnov two-sample test and Pearson's chi-squared test [132], and for the independence problem: Pearson's correlation (e.g. [140]) and Spearman's rank correlation coefficient (e.g. [193]), among others. With increasingly heterogeneous data collection practices and driven by a need to handle more complex data types, a range of recent nonparametric tests have been developed, applicable to multivarite and structured data, including tests based on distances in reproducing Kernel Hilbert spaces [71, 69, 70], mutual information criteria using permutation techniques [18] and using permutation techniques more generally [19]. Deep learning has also emerged as an alternative for defining tests on structured objects. [114] define classifier two-sample tests and [108] use deep kernels to embed structured objects. Tests in most of these cases, however, are defined directly on the space of observations, it is not clear how to input examples of varying sizes, or how to account for the uncertainty in individual observations especially if these change across sets.

In the context of kernel methods, note that kernels defined on sets directly [95], measuring the similarity between sets by the average pairwise point similarities between the sets, are not known to be characteristic. Attempts have also been made to define kernels on the space of distributions, including probability product kernel [84], the Fisher kernel [81], diffusion kernels [98] and kernels arising from Kullback-Leibler divergences [119], none of them known to be characteristic and in this case with the shortcoming that many of the above are parameterized by a family of densities which may or may not hold in data.

We make a note that accommodating for input uncertainty has connections with robust hypothesis testing. These tests attempt to explicitly enforce invariances in test statistics in a certain uncertainty ball to remove irrelevant sources of variation [62, 75]. Other types of invariances can also be enforced, for instance [101] use features designed to be invariant to additive noise

and use distances between those representations for hypothesis testing. One may also use a model-based approach to capture this uncertainty, for instance [15] use Gaussian processes and compare posterior distributions. More generally, also work in the functional data analysis literature [196, 126] uses a model-based approach to testing sets that represent functions.

2.2 Hypothesis Tests on Sets

In the following sections, we propose tests to evaluate two common hypotheses: the two sample problem of testing equality of distributions in two samples, and the independence problem of testing whether joint distributions in paired samples coincide with the product of their marginals.

For both tests, the exposition mirrors well-known results in kernel hypothesis testing which we will only briefly describe (see [70, 73] for more background). The contribution of this chapter is to show that tests defined with a second level of sampling are consistent and to show that correctly weighting representations according to their set size is most efficient.

Algorithm. We may summarize hypothesis testing in this context as follows:

1. Embed the distributions $\{\mathbb{P}_i\}_{i=1}^N$ into an RKHS using approximations of the mean embeddings $\{\hat{\mu}_{\mathbb{P}_i}\}_{i=1}^N$ computed with independent samples $\{x_{i,j}\}_{j=1}^{n_i} \sim \mathbb{P}_i$.
2. Define test statistics on this feature representations to test for a certain hypothesis or dependency structure in $\mathcal{M}$.

2.2.1 The two sample problem

Consider a first collection of sets of observations, each i-th set denoted $\{x_{i,s}\}_{s=1}^{n_i} \sim \mathbb{P}_i$, for a total of N such sets with distributions $\{\mathbb{P}_i\}_{i=1}^N \sim \mathcal{M}_P$, and define similarly a second collection of sets, each j-th set $\{y_{j,s}\}_{s=1}^{n_j} \sim \mathbb{Q}_j$, for $\{\mathbb{Q}_j\}_{j=1}^M \sim \mathcal{M}_Q$. The problem we consider is to test whether,

$$\mathcal{H}_0 : \mathcal{M}_P = \mathcal{M}_Q \quad \text{or else} \quad \mathcal{H}_1 : \mathcal{M}_P \neq \mathcal{M}_Q, \tag{2.3}$$

holds on the basis of the observations available in each set. We illustrate this problem in Figure 2.1. The proposed test statistic approximates the square of the RKHS distance between densities

$\mathcal{M}_P$ and $\mathcal{M}_Q$, also called Maximum Mean Discrepancy (MMD), which may be decomposed as follows [70],

$$\mathrm{MMD}^2 := \mathbb{E}_{\mathbb{P},\mathbb{P}'\sim\mathcal{M}_P} K(\mathbb{P},\mathbb{P}') + \mathbb{E}_{\mathbb{Q},\mathbb{Q}'\sim\mathcal{M}_Q} K(\mathbb{Q},\mathbb{Q}') - 2\mathbb{E}_{\mathbb{P}\sim\mathcal{M}_P,\mathbb{Q}\sim\mathcal{M}_Q} K(\mathbb{P},\mathbb{Q}), \tag{2.4}$$

where K is the kernel on distributions given after equation (2.2). We denote $\widehat{\mathrm{MMD}}^2$ the empirical estimator of the MMD^2 with expectations replaced by averages, obtained from independent samples $\{\mathbb{P}_i\}_{i=1}^N \sim \mathcal{M}_P$ and $\{\mathbb{Q}_j\}_{j=1}^M \sim \mathcal{M}_Q$. The proposed statistic is defined by considering approximate mean embeddings of each distribution and considering the weighted sample of their meta-distribution each of them represents,

$$\widehat{\mathrm{RMMD}}^2 := \sum_{i,j=1}^{N} w_{\mathbb{P}_i} w_{\mathbb{P}_j} K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j}) + \sum_{i,j=1}^{M} w_{\mathbb{Q}_i} w_{\mathbb{Q}_j} K(\hat{\mu}_{\mathbb{Q}_i}, \hat{\mu}_{\mathbb{Q}_j}) - 2 \sum_{i,j=1}^{N,M} w_{\mathbb{P}_i} w_{\mathbb{Q}_j} K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{Q}_j}).$$

R stands for robust. Assume for now that all weights are fixed $w_{\mathbb{P}_i} = 1/N, w_{\mathbb{Q}_j} = 1/M$ for all i, j. We return to the specification of weights in section 2.2.3. The asymptotic behaviour of $\widehat{\mathrm{MMD}}^2$ is well understood [70] and the test itself is extensively used in many applications [109, 137]. However, these results do not extend trivially if each independent set exhibits an additional source of variation due to the estimation of the mean embedding. In the following proposition, we bound the contribution of this additional source of variation and show that under the asymptotic regime where both the set sizes and number of sets grow larger, asymptotic distributions are well defined.

Proposition 1 ***(Asymptotic distribution)***. *Let two samples of data be defined as above and let K be characteristic and L_K-Lipschitz continuous. Then, under the null and alternative and in the regime of increasing set size n_i and increasing sample sizes N and M, the asymptotic distributions of $\widehat{RMMD}^2$ coincides with that of $\widehat{MMD}^2$.*

Proof. All proofs are given in the Appendix.

In other words, the additional variability due to a second level of sampling converges to 0 asymptotically, and thus the asymptotic distribution converges to that of the well known MMD two sample test of [70].

2.2.2 The independence problem

Independence tests are concerned with the question of whether two random variables are distributed independently of each other. For this problem, we start with a collection of *paired*

distributions $\{(\mathbb{P}_i, \mathbb{Q}_i)\}_{i=1}^N$ drawn from a joint distribution we denote $\mathcal{M}_{PQ}$, and denote their marginals $\mathcal{M}_P$ and $\mathcal{M}_Q$. The hypothesis problem is to determine whether,

$$\mathcal{H}_0 : \mathcal{M}_{PQ} = \mathcal{M}_P \mathcal{M}_Q \quad \text{or else} \quad \mathcal{H}_1 : \mathcal{M}_{PQ} \neq \mathcal{M}_P \mathcal{M}_Q. \tag{2.5}$$

Example. Consider an example from healthcare to illustrate this problem.

- A similar set-up as that given in Figure 2.1 may be used to illustrate independence testing with set-valued data. A common problem is identify dependencies between biomarkers, often observed irregularly over time in many patients. For instance cholesterol levels $\{x_{i,t_1}, \ldots, x_{i,t_{n_i}}\}$ and blood pressure $\{y_{i,t_1}, \ldots, y_{i,t_{n_i}}\}$ may be observed over times $t_1, \ldots, t_{n_i}$ in N individuals $i = 1, \ldots, N$. To formally test for dependencies between these samples one must account for the irregularity in observation time and uncertainty in biomarker reads. This can be done by considering instead distributions $\mathbb{P}_i$ and $\mathbb{Q}_i$ and testing for independence in this space directly.

As in the two-sample test, we may quantify the difference between distributions using the RKHS distance $||\mu_{\mathcal{M}_{PQ}} - \mu_{\mathcal{M}_P} \otimes \mu_{\mathcal{M}_Q}||^2_{HS}$. Kernels K, L are assumed characteristic; $||\cdot||_{HS}$ is the norm on the space of $\mathcal{H}_K \to \mathcal{H}_L$ Hilbert-Schmidt operators, and $\otimes$ denotes the tensor product, such that $(a \otimes b)c = a\langle b, c\rangle$. This distance is called the Hilbert Schmidt Independence Criterion (HSIC) [71, 73].

Two empirical estimators can be written: one assuming access to independent samples $\mathcal{M}_{PQ}$ and one with independent samples from each of the paired distributions sampled from $\mathcal{M}_{PQ}$,

$$\widehat{\text{HSIC}} = \text{Tr}\,(KHLH)/N^2, \qquad \widehat{\text{RHSIC}} = \text{Tr}\,(\hat{K}H\hat{L}H) \cdot N^2, \tag{2.6}$$

for kernel matrices with (i, j) entries $K_{ij} = K(\mathbb{P}_i, \mathbb{P}_j) = \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle_{\mathcal{H}_K}$ and $L_{ij} = \langle \mu_{\mathbb{Q}_i}, \mu_{\mathbb{Q}_j} \rangle_{\mathcal{H}_L}$ for the population version and $\hat{K}_{ij} = w_{\mathbb{P}_i} w_{\mathbb{P}_j} \langle \hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j} \rangle_{\mathcal{H}_K}$ and $\hat{L}_{ij} = w_{\mathbb{Q}_i} w_{\mathbb{Q}_j} \langle \hat{\mu}_{\mathbb{Q}_i}, \hat{\mu}_{\mathbb{Q}_j} \rangle_{\mathcal{H}_L}$ with mean embeddings replaced by their weighted finite sample counterparts for the robust alternative. Assume for now that all weights are fixed $w_{\mathbb{P}_i} = 1/N, w_{\mathbb{Q}_j} = 1/M$ for all i, j. The centering matrix is defined by $H = I - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ and Tr is the trace operator.

Here, similarly to the two sample problem, approximations due to a second level of sampling are well behaved and mirror those of the robust statistic for the two-sample problem. In particular, that asymptotic distributions of the RHSIC and the HSIC coincide in the regime with increasing set size and increasing sample size, making hypothesis testing with the $\widehat{\text{RHSIC}}$ consistent for the independence problem in equation (2.5).

Proposition 2 ***(Asymptotic distribution)***. *Let two samples of data be defined as above and let K be characteristic and L_K-Lipschitz continuous. Then, under the null and alternative and in the regime of increasing set size n_i and increasing sample size N, the asymptotic distributions of $\widehat{RHSIC}$ coincides with that of $\widehat{HSIC}$.*

Independence testing with the $\widehat{\text{HSIC}}$ has been studied in [73, 198, 85].

2.2.3 Practical remarks

We make a number of remarks on the practical application of our tests.

- **Weights for high power.** Set sizes in practice may be limited. In the asymptotic regime of increasing number of sets but finite set size, the properties of the estimator may depend on appropriately weighting sets for high power. The proposed weighting scheme addresses this point.

  Recall that each individual observation x_{ij} is drawn independently from their respective distributions $\mathbb{P}_i$. Other factors of variations assumed to be common across sets, the variance of the approximate embedding $\hat{\mu}_{\mathbb{P}_i}$ is therefore proportional to $1/n_i$ (i.e. the variation in approximation of mean embeddings is due solely to diverging set sizes). When mean embeddings have different variances, it is efficient to give less weight to mean embeddings that have high variances. By efficient in this context, we mean highest asymptotic power of tests based on mean embedding representations of sets.

  For V-statistics the asymptotic power function is well known, and an argument involving the delta method for differentiable kernels, expanded on in the Appendix, can be used to determine the optimal weights to be given by $w_{\mathbb{P}_i} := n_i / \sum_i n_i$ for each i.

- **Hyperparameters for high power.** With a similar intuition, even though in theory we can expect high power for any alternative hypothesis and any choice of kernel, with finite sample size, some kernel hyperparameters will give higher power than others. The proposed tests optimize the choice of kernels by choosing hyperparameters that minimize the asymptotic variance under the alternative similarly to [169, 85]. But, in addition, we extend the optimization to tune both the mean embedding to represent sets and the kernel used for comparisons in Hilbert space. Please find more details in the Appendix.

- **Low-dimensional approximations for large scale data.** Testing on distributions as described is often not scalable for even to large datasets, as computing each of the entries of the relevant kernel matrices requires defining a high-dimensional mean embedding. To define

test statistics on these representations we further embed the non-linear feature space $\mathcal{H}_k$ defined by k into a random low dimensional Euclidean space using their expansion in Hilbert space as a linear combination of the Fourier basis [146, 136]. If we draw m samples from the Gaussian spectral measure, we can approximate the Gaussian kernel k by,

$$k(x,y) \approx \frac{2}{m}\sum_{j=1}^{m}\cos(\langle\omega_j,x\rangle+b_j)\cos(\langle\omega_j,y\rangle+b_j) = \langle\phi(x),\phi(y)\rangle,$$

where $\omega_1,\ldots,\omega_m \sim \mathcal{N}(0,\gamma)$, $b_1,...,b_m \sim \mathcal{U}[0,2\pi]$, and $\phi(x) = \sqrt{\frac{2}{m}}[\cos(\omega_1 x+b_1),...,\cos(\omega_m x+b_m)] \in \mathbb{R}^m$ [136]. The mean embedding $\mu_{\mathbb{P}} = \mathbb{E}_{X\sim\mathbb{P}}\phi(X)$ can then be approximated with elements in the span of $(\cos(\langle\omega_j,x\rangle+b_j))_{j=1}^m$. By averaging over the available n_i samples in X_i from the distribution $\mathbb{P}_i$, the approximate finite-dimensional embedding is given by,

$$\hat{\mu}_{\mathbb{P}_i,m} = \frac{1}{n_i}\sum_{x\in\{x_{ij}\}_{j=1}^{n_i}}\sqrt{\frac{2}{m}}(\cos(\langle w_j,x\rangle+b_j))_{j=1}^m \in \mathbb{R}^m.$$

2.3 Synthetic Data Experiments

The purpose of synthetic experiments will be to test **power**: the rate at which we correctly reject $\mathcal{H}_0$ when it is false, as we increase the difficulty of the testing problems; and **Type I error**: the rate at which we incorrectly reject $\mathcal{H}_0$ when it is true.

In all experiments, α (the target Type I error) is set to 0.05, the number of time series is set to $N = 500$, the number of observations made on each time series is random between 5 and 50, and each problem is repeated for 500 trials.

Tests for empirical comparisons. To the best of our knowledge, no existing test naturally accommodates for set-valued data with irregular sizes. Our approach to empirical comparisons will be to coerce the data into a fixed dimensional vector in a well-defined manner, and evaluate existing tests on this representation. To do so, we focus on time-series -like data which we interpolate along the time axis with cubic splines and evaluate at a fixed number of time points.

- The following tests are evaluated for the two-sample problem. The **MMD** [70] with hyper-parameters optimized for maximum power, two-sample classifier tests [114] which involve fitting a deep classifier. We considered a recurrent neural network with GRU cells for sequential data (**C2ST-GRU**) and the DeepSets approach of [192] modelling permutation

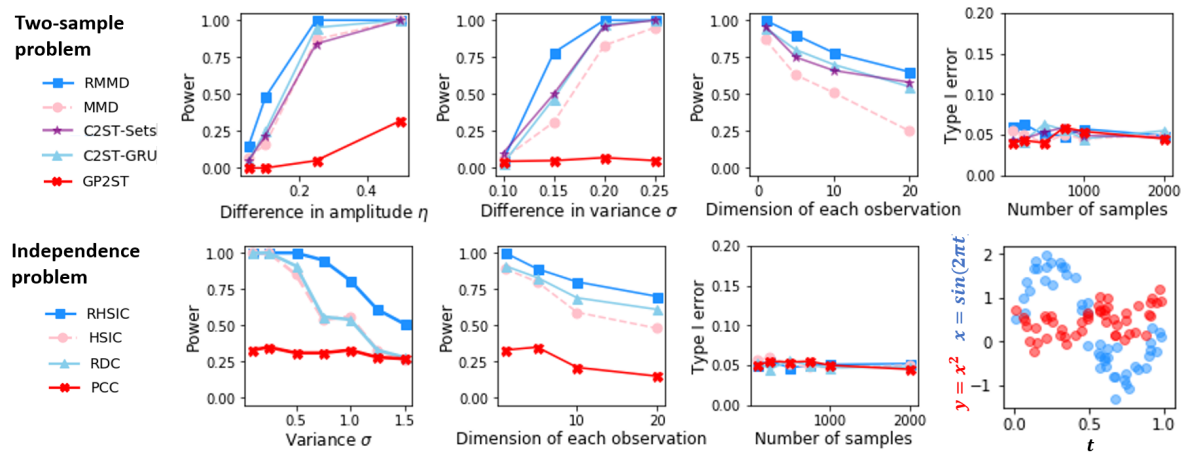


Figure 2.2: Power (higher better) and Type I error (at level 0.05) on synthetic data. The rightmost panel gives type I error with approximate control at the level $\alpha = 0.05$ for all methods. **Top row**: two-sample problem. **Bottom row**: independence problem. RMMD and RHSIC are the proposed tests.

invariance to be expected in sets (**C2ST-Sets**). We consider also the Gaussian process-based test (**GP2ST**) by [15].

- For the independence problem we consider: the **HSIC** [73], the Randomized Dependence Coefficient (**RDC**) [112] and Pearson Correlation Coefficient (**PCC**).

For all kernel-based tests, because their null distributions are given by an infinite sum of weighted χ^2 variables (no closed-form quantiles), in each trial we use 400 random permutations to approximate the null distribution. All independence tests also use 400 random permutations to approximate the null distribution. C2ST-based tests uses its asymptotic distributions under the null for significance thresholds and GP2ST uses credible intervals. We give more details on the implementation of each of these tests in the Appendix.

2.3.1 Two-sample problem

Experiment design. Each one of the two samples is defined by a family of N distributions $\{\mathbb{P}_i\}_{i=1}^N$ we take to be Gaussian $\mathbb{P}_i = \eta \sin(2\pi t) + \mathcal{N}(0, \sigma_i + \sigma)$. The variability between the $\{\mathbb{P}_i\}_{i=1}^N$ is specified by σ_i, drawn from a one-parameter inverse gamma distribution, which mimics the behaviour of the meta-distribution and the observation pattern we may observe in heterogeneous data. The difference between two populations of sampled distributions is the mean amplitude η and/or shifts in *baseline* variance σ.

Two-sample problems become harder whenever these parameters converge to the same value in the two samples and are easier when they diverge. The sampled Gaussian distributions themselves are not observable and, in turn, we have access to observations $x_{ij} \sim \mathbb{P}_i$. Each x_{ij} is obtained by fixing t to $t_j \sim \mathcal{U}[0,1]$ and subsequently sampling from the Gaussian.

The result is two collections of noisy time series with non-linear dynamics. Each time series, or set of observations, is irregularly sampled with noise levels that vary between sets.

Results. We report performance for the two sample problems in the **top row** of Figure 2.2. Power is measured in three experiments: *first*, as we increase the difference in time series amplitude (with equal variance $\sigma = 0.1$), second as we increase the observation variance (with equal amplitude $\eta = 1$) between the two populations, and third as the dimension of each time series increases (on data sampled with a single dimension with a difference in amplitude equal to 0.25 and other dimensions with no difference). Type I error is shown as a function of the number of samples.

All tests approximately control for type I error at the desired threshold. In terms of power, we observe the RMMD to outperform across all experiments with an important contrast on the difference in performance with the MMD. Even though using similar test statistics, the RMMD much more faithfully captures the irregularity and uncertainty of every individual set of observations. RMMD similarly outperforms C2ST-based tests, the strongest baselines, with up to a two-fold increase in power for small differences in amplitude and variance.

2.3.2 Independence problem

Experiment design. We aim to construct pairs of distributions $(\mathbb{P}_i, \mathbb{Q}_i)$. Define the mean of each distribution $\mathbb{P}_i$ as $f_i(t) := \beta_i \sin(2\pi t) + \alpha_i t$. Differently than in the two-sample problem, the variability among the $\{\mathbb{P}_i\}$ appears in the amplitude and trend of the sine function, let these be $\beta_i \sim \mathcal{U}[0.5, 1.5]$ and $\alpha_i \sim \mathcal{U}[-0.5, 0.5]$. Once these parameters are sampled, paired distributions $(\mathbb{P}_i, \mathbb{Q}_i)$ are given by $\mathbb{P}_i = f_i(t) + \mathcal{N}(0, \sigma)$ and $\mathbb{Q}_i = g(f_i(t)) + \mathcal{N}(0, \sigma)$. Each observation from this pair is obtained as in the two sample problem by fixing a random t and sampling from the resulting distribution.

The difficulty of the problem is governed by two factors: g and σ. g determines the dependency between the two functions. In every trial, $g(x)$ is randomly chosen from the set of functions $\{x^2, x^3, \cos(x), \exp(-x)\}$. Testing for dependency is hard also for increasing variance σ of observations, as this makes the dependent paired samples appear independent. A sample of

dependent sets of data using this data generating mechanism is given in the lower rightmost panel of Figure 2.2.

Results. Power and type I error are shown in the bottom row of Figure . The bottom row of Figure 2.2 gives performance results for the independence problem. In the first two leftmost panels we evaluate power as we increase the variance of paired time series and as we increase the dimensionality of each observation for a fixed variance $\sigma = 0.5$. The bottom rightmost plot shows a sample of two dependent noisy time series, colored blue and red respectively, for illustration.

The conclusions for this problem mirror the two-sample testing experiments, with however a much larger increase in power over alternatives, all using less flexible data representations as none of them avoids interpolating between observations before testing independence which we hypothesize is one reason for their underperformance. This is consistent with the increasing variance experiment, in this case increasing variance worsens interpolation performance.

2.4 Testing on lung function data of Cystic Fibrosis patients

For people with Cystic Fibrosis (CF), mucus in the lungs is linked with chronic infections that can cause permanent damage, making it harder to breathe [90]. This condition is often measured over time using `FEV1% predicted`; the Forced Expiratory Volume of air in the first second of a forced exhaled breath we would expect for a person without CF of the same age, gender, height, and ethnicity [174]. For example, a person with CF who has `FEV1% predicted` equal to 50% can breathe out half the amount of air as we would expect from a comparable person without CF. In this experiment, we work with data from the UK Cystic Fibrosis Trust containing records from $10,980$ patients with approximately annual follow ups between 2008 and 2015, with the objective of better understanding the dependence of lung function over time with other biomarkers. For this problem we found a significant influence of Body Mass Index (BMI) over time and the number of days under intravenous antibiotics in a given year; both already known to be associated with lung function [185, 91].

We use this information to create a set of problems under the alternative $\mathcal{H}_1$ with an additional twist. We increase heterogeneity among patients by artificially removing a proportion p of densely sampled patients (here more than 4 recordings). The problem is to test for independence between a patients two-dimensional trajectory of BMI and antibiotics measurements over time, and their lung function trajectory over time. In this set-up, we expect the information content of

the average patient to decrease, a scenario that lends itself to an importance-weighted approach (more weight on densely sampled trajectories), such as described in section 2.2.3. In this section we test this property, which we found advantageous for higher missingness data patterns, as shown in Figure 2.3. In this case, power tends to be higher after weighting (RHSIC) versus not weighting (RHSIC-weight). We report also type I errors, well controlled by all methods, evaluated after shuffling the lung function trajectories between patients, such as to break the associations between BMI and antibiotics, and lung function trajectories.

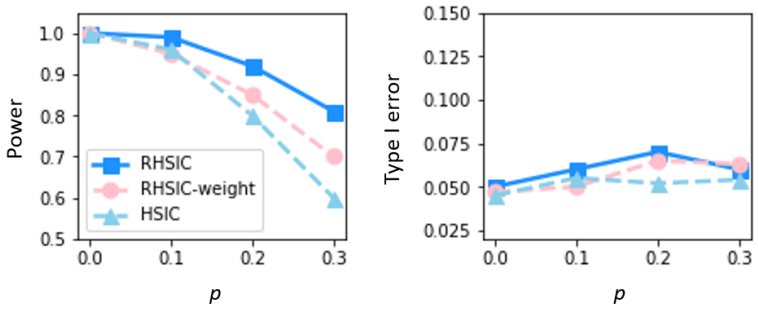


Figure 2.3: Power and Type I error on Cystic Fibrosis data.

2.5 Testing on Climate Data

This experiment explores the use of extensive weather data to determine whether the recent rapid changes in climate associated with human-induced activities significantly differ from natural climate variability. A number of variables are used to monitor the state of the climate including precipitation, wind patterns, and atmospheric composition among others. It depends on the latitude and longitude, and regions may vary and evolve differently over time [166].

Interpretation as set-valued data. We can think of the multivariate measurements in different locations across the globe at a given time as a set of data points. Each set sampled from a probability distribution that represents the global weather pattern of the climate. We follow standard descriptions to define the climate as a collection of these sets observed over a period of 20 years. The problem is to test for significant differences in climate, represented by the evolution of bags of (multi-channel) images, over time (see Figure 2.4).

Experiment design. The data is publicly available, provided by the Copernicus Climate Change Service[2]. We include a total of 12 climate variables identified as essential to characterize

[2]https://climate.copernicus.eu/.

Figure 2.4: Illustration of the two-sample problem with *global* set-valued data versus *local* time series data.

the climate[3], including temperature, atmospheric pressure, observed over monthly periods for the last 40 years across Europe. The available data thus consists of a two streams of sets $\{x_{i,j}\}_{j=1}^{n_i}$ and $\{y_{i,j}\}_{j=1}^{n_i}$ for $i = 1, \ldots, 144$ (12 months over 20 years). The first describes the climate over the period $1979-1999$, and the second set over the period $1999-2019$. Both contain measurements $x_{i,j} \in \mathbb{R}^{12}$ ($y_{i,j}$ respectively) in *approximately* $n_i = 250$ different locations (approximately because not all locations are consistently observed over time) which makes the length of each set irregular. Existing tests would thus require some form of interpolation which is not trivial over space and time in this case.

Problem. The problem is to test for the hypothesis of equally distributed climate data over the past 4 decades. We make different test: on data from the European, African, North American, South American and South-East Asian regions.

Results. RMMD rejects the hypothesis of equally distributed climate data over the past 4 decades in Europe (p-value 0.0002), Africa (p-value 0.0014), and South America (p-value 0.0001) but fails to reject at a level of 0.01 for North America (p-value 0.016) and South-East Asia (p-value 0.036). In the case of Europe, we note that this result would be different if only a particular location was considered (which could have been a viable reductionist strategy to use existing tests). For instance, we found that the RMMD applied to climate data over the same periods in London and Paris to not be significantly different (p-value 0.21). This experiment demonstrates the potential benefits of using more flexible tests that better represent available data to faithfully investigate complex phenomena such as climate that involve multiple measurements over time and space.

[3] https://public.wmo.int/en/programmes/global-climate-observing-system/essential-climate-variables

2.6 Conclusions

In this chapter, we extended the toolkit of applied statisticians to do hypothesis testing on *set*-valued data. We have shown that by appropriately representing each set of observations in a Hilbert space, kernel-based hypothesis testing may be applied consistently. Specifically, we introduced tests for the two-sample and the independence problem, derived their asymptotic distributions and provided efficient algorithms and optimization schemes to analyse a wide range of scenarios in an automatic fashion.

Chapter 3

A Kernel Two-Sample Test with Sample Selection Bias

Hypothesis testing can help decision-making by quantifying distributional differences between two populations from observational data. However, these tests may inherit biases embedded in the data collection mechanism (some instances often being systematically more likely included in our sample) and consistently reproduce biased decisions. In this chapter, we propose a two-sample test that adjusts for selection bias by accounting for differences in marginal distributions of confounding variables. Our test statistic is a weighted distance between samples embedded in a reproducing kernel Hilbert space, whose balancing weights provably correct for bias. We establish the asymptotic distributions under null and alternative hypotheses, and prove the consistency of empirical approximations to the underlying population quantity. We conclude with performance evaluations on artificial data and experiments on treatment effect studies from economics.

Introduction

The two-sample problem considers testing whether two independent samples are likely drawn from the same distribution. Such tests have a long history in statistical inference but they are also increasingly used in decision making scenarios. For example, two-sample tests have been used to determine gender differences in academic achievements [80], gender differences in criminal justice outcomes [68], gender differences in health issues [181], and also frequently used in medicine to determine subgroups of patients that respond differently to medication and establish treatment policies [22].

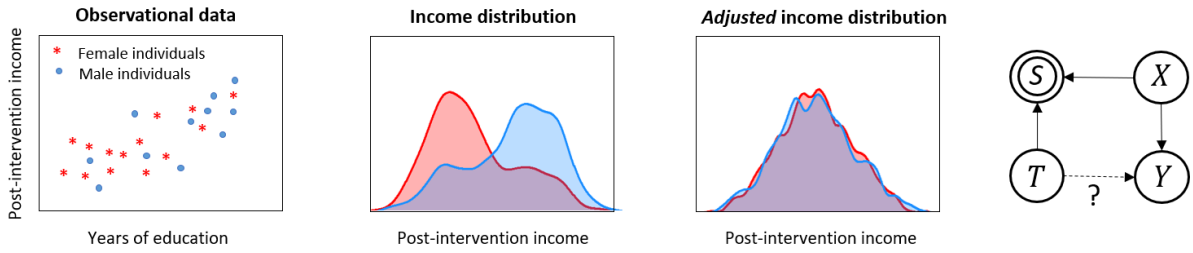


Figure 3.1: **The influence of selection bias**. The left panel plots a sample from the observed data, the middle panel shows the observed post-intervention income density for males and females, while the right panel shows the income distribution obtained by adjusting for education levels (we partition into homogeneous groups before aggregating their densities, see a description of the problem in the introduction and details of the data generating mechanism in Appendix B.2). In this case, a conventional two-sample test rejects the hypothesis of equal post-intervention income in male and female populations due to the intervention, while our proposed test fails to reject.

In any data driven study, a *first* step is the collection of a series of observations about an underlying phenomenon of interest before making an informed decision, for example assisted by a hypothesis test, on this data. In most realistic scenarios, we do not have control on the data collection process (e.g. participants volunteering for a study involving a new treatment may differ systematically from the wider population), but we do implicitly condition on the fact that participants entered into the study $(S = 1)$.

This implicit conditioning may *bias* the conclusions of tests because two samples may differ systematically prior to running any experiment and a hypothetical difference in distribution be completely unrelated to the effect of interest.

The problem of selection bias and its influence on inference has attracted much recent interest in the fairness literature [130, 88, 94, 51], one aspect of which involves mitigating indirect discrimination e.g., section 3.1, point (2) in [204], in which algorithms make biased decisions due to the correlation of the non-discriminatory items with the discriminatory ones. Selection bias is also relevant in the causal inference literature [128]. [8] gave graphical conditions under which the causal effect may be recovered from data with selection bias. A similar scenario is considered under the rubric of treatment effect estimation, in which algorithms estimate individualized, average and conditional treatment effects in data biased by confounders that simultaneously influence treatment assignment and outcomes [186, 87, 199]. In epidemiology [141] and econometrics [77], versions of this problem are also widely studied. Similarly, hypothesis test for the significance of a measured association, and data-driven algorithms in general, must account for sources of discrimination, confounding, and selection bias more generally in the data.

In fairness and causal inference however, while many methods exist attempting to predict associations adjusting for selection bias, much less is known on the *significance* of effects in the presence of selection bias. We cannot for instance say whether outcome distributions in two groups *significantly* differ or not even if model predictions differ. The literature on hypothesis testing is invested in such problems but so far hypothesis testing in the presence of selection bias has been minimally considered. To illustrate the large impact of selection bias on two-sample testing outcomes and the need for approaches that can adjust for these spurious effects we consider an example described below and illustrated in Figure 3.1.

Example. Suppose a city government wants to understand the role of gender on the effectiveness of a past employment program to better allocate their resources in the future. Its analyst constructed datasets of volunteering ($S = 1$) men and women ($T = 1$ and $T = 0$) to be compared, and included a number of relevant employment figures such as post-intervention earnings, type of job, satisfaction, etc. (Y). In this hypothetical example, highly educated men were more likely to volunteer than women due to historical gender bias in education opportunities (X). Such preferential selection creates a *spurious* association between T and Y, opening a path of unblocked correlations through X, as shown in the causal diagram of Figure 3.1. It is called spurious because it is not part of what we seek to estimate – **the significance of the causal effect of T on Y**.

A test that ignores this bias tends to determine men and women to have different employment program outcomes whereas in reality, once we account for differences in education (i.e. we block the spurious open path), the program is found to perform equally in distribution across men and women. In this example, higher program benefits are due to higher starting education standards, not because people of different gender benefit differently. A decision based on a plain two sample tests overrates the impact of an individual's sex – in this case correlated with education because we implicitly condition on $S = 1$. Please find a description of the data generating mechanism in the Appendix.

Contributions. We develop a non-parametric test for differences in distribution of two samples biased by preferential selection driven by other observed quantities. Our proposal is a generalization of two sample tests based on maximum mean discrepancies between probability distributions [70, 39, 86, 194, 12] that incorporate importance sampling techniques to adjust for distributional shift in covariates. The technical challenge is that adjustments made for differences in the marginal confounding distributions between two samples are data-dependent, and therefore invalidate existing asymptotic guarantees of tests based on the maximum mean discrepancy.

Our contributions are three-fold.

1. We propose a two-sample test statistic that, under certain conditions, provably adjusts for selection bias.
2. We derive novel asymptotic distributions for the proposed test.
3. In the finite-sample case, we propose weight approximations for our test statistic, that we show to be consistent with its population-level quantity.

3.1 Background

From the context of hypothesis testing, to understand the role of selection bias it is useful to bring in knowledge of the causal mechanisms in data and augment a causal graph with a variable S that represents the recruitment of individuals into the study. The assignment of individuals into two groups $T \in \{0,1\}$ is then correlated with confounding variables $X \in \mathcal{X}$ through the fact that we condition on individuals to be included in the study (see Figure 3.1). We call these confounding variables because they introduce spurious differences in the relationship between outcome variables and the selection mechanism once we condition on $S = 1$. To formalise hypothesis testing with biased data, we adopt the potential outcomes framework of [145]. We assume to have observed independent samples from and outcome variable $Y = Y^1 \cdot T + Y^0 \cdot (1-T)$, the response variable Y is split into counterfactual variables, Y^0 and Y^1, which appeal to the potential values of an individual were $T = 0$ and $T = 1$ respectively, i.e. under a model where selection bias does not influence treatment assignment.

The hypothesis testing problem is formulated as evaluating the evidence for a difference in distribution P_{Y^1} and P_{Y^0} in two groups of observations,

$$\mathcal{H}_0 : P_{Y^1} = P_{Y^0} \quad \text{versus} \quad \mathcal{H}_1 : P_{Y^1} \neq P_{Y^0}, \tag{3.1}$$

but, unlike conventional two-sample problems, we have access to distributions P_{Y^1} and P_{Y^0} only via an (unknown) sampling policy $T \in \{0,1\}$ that introduces bias due to the implicit conditioning on $S = 1$, rather than directly through independent samples from P_{Y^1} and P_{Y^0}. S and T create distributional shift, the assumption is that the available data is independently sampled from *distorted* distributions conditional on T. The counterfactual distributions P_{Y^0} and P_{Y^1} we are interested in differentiating are not directly observed and instead through available samples we have access to $P_{Y|T=0}$ and $P_{Y|T=1}$, different from P_{Y^0} and P_{Y^1} because $(Y^1, Y^0) \not\perp\!\!\!\perp T | S = 1$.

The same attributes X that correlate with the probability of group assignment T may also be associated with the potential responses Y^0 and Y^1.

3.1.1 Preliminaries on Hypothesis Testing

The problem of hypothesis testing is to define a test statistic (a function of observational data) to distinguish between two hypotheses on the distribution of observed samples. Short of perfectly distinguishing between any two hypotheses we may pose due to the limited number of samples available to characterize distributions, tests are constructed such that a certain hypothesis is rejected whenever a test statistic exceeds a certain threshold away from 0 [103]. The goal of hypothesis testing is to derive a threshold such that false positives are upper bounded by a design parameter α and false negatives are as low as possible.

Our test statistic is characterized by distances in mean embeddings of distributions in a Reproducing kernel Hilbert space $\mathcal{H}_k$. The advantage of mapping distributions P_{Y^0} and P_{Y^1} to functions in $\mathcal{H}_k$ is that we may now say that P_{Y^0} and P_{Y^1} are close if the RKHS distance $||\mu_{P_{Y^0}} - \mu_{P_{Y^1}}||_{\mathcal{H}_k}$ is small, where $\mu_P := \int_{\mathcal{X}} k(x,\cdot)dP(x)$ is the embedding of the probability measure P to $\mathcal{H}_k$. This distance is known as the Maximum Mean Discrepancy (MMD) [70] and is particularly appealing because for certain choices of the kernel function k, the mean embedding can be shown to be injective [165]. All properties of the distribution are conserved with this map and one may distinguish between distributions by computing the MMD between them.

$$\text{MMD}(P_{Y^0}, P_{Y^1}) = 0 \quad \text{if and only if} \quad P_{Y^0} = P_{Y^1}, \tag{3.2}$$

We focus our attention on the Gaussian kernel $k(x,y) = \exp(-||x-y||^2/\sigma^2)$ with bandwidth parameter σ, that enjoys this property. The squared MMD is given by [70],

$$\text{MMD}^2 := \mathop{\mathbb{E}}_{y,y^\star \sim P_{Y^1}} k(y,y^\star) + \mathop{\mathbb{E}}_{y,y^\star \sim P_{Y^0}} k(y,y^\star) - 2 \mathop{\mathbb{E}}_{y \sim P_{Y^1}, y^\star \sim P_{Y^0}} k(y,y^\star), \tag{3.3}$$

and empirical estimates may be computed in practice.

3.2 An Importance Weighted Statistic

With access only to samples from biased populations $P_{Y|T=1}$ and $P_{Y|T=0}$ estimating the above distance with respect to counterfactual distributions P_{Y^0} and P_{Y^1} empirically is not possible. To ensure identifiability of the hypothesis testing problem however, we may assume that (Y^0, Y^1) and the data generating process satisfy ignorability: $Y^0, Y^1 \perp\!\!\!\perp T|X, S=1$, a common assumption in the treatment effect estimation literature. It means that within any stratum of X, individuals who would have one set of potential outcomes $Y(0)=y_0$ and $Y(1)=y_1$, are just as likely to be in the control or treatment group as other individuals (with different potential outcomes) that share characteristics X. If in addition we assume that $0 < Pr(T|X) < 1$, then with knowledge of the sample selection mechanisms $e(x) := Pr(T=1|X=x)$ we may recover the expectations of interest with importance sampling,

$$\mathbb{E}\left(\frac{Y}{e(X)}\,\middle|\, T=1\right) = \mathbb{E}\left(\frac{T\cdot Y^1}{e(X)}\right) = \mathbb{E}\left(\mathbb{E}\left(\frac{T\cdot Y^1}{e(X)}\,\middle|\, X\right)\right) = \mathbb{E}\left(Y^1\right). \tag{3.4}$$

This encourages us to define a weighted estimator of the MMD - called the WMMD - such that the weights emphasize distances in areas of the support where the distributions of confounding variables agree. Define w such that $Pr(T=1|X=x)\cdot w(x) = Pr(T=0|X=x)$ and consider,

$$\text{WMMD}^2 := \mathop{\mathbb{E}}_{P_{XY|T=1}} w(x)w(x^\star)k(y,y^\star) + \mathop{\mathbb{E}}_{P_{Y|T=0}} k(y,y^\star) - 2\mathop{\mathbb{E}}_{\substack{x,y\sim P_{XY|T=1},\\ y^\star\sim P_{Y|T=0}}} w(x)k(y,y^\star), \tag{3.5}$$

where the superscript $\star$ denotes an independent copy where appropriate. We show next that this metric consistently distinguishes between null and alternative hypotheses at the population level.

Proposition 1 *For k a characteristic kernel and known weights $w(x) > 0$ for all $x \in \mathcal{X}$, WMMD $= 0$ if and only if $P_{Y_1} = P_{Y^0}$.*

Proof. All proofs are given in the Appendix.

A kernel k is characteristic if the mean embedding μ_P is injective [70]. In practice, we have access to an empirical estimate of the WMMD, defined as follows,

$$\widehat{\text{WMMD}}^2 := \sum_{i\neq j: t_i=t_j=1} w(x_i)w(x_j)k(y_i,y_j) + \sum_{i\neq j: t_i=t_j=0} k(y_i,y_j) - 2\sum_{i,j: t_i=1, t_j=0} w(x_i)k(y_i,y_j),$$

where the (y_i, t_i, x_i) are realizations of the random variables (Y, T, X). Deviations from 0 (the theoretical value under the null) are expected due to finite sample variation. Tests are then

constructed such that the null hypothesis is rejected whenever $\widehat{\text{WMMD}}^2$ exceeds a certain threshold. In the next section we will show how to consistently define such a threshold to ensure a low margin of error.

3.2.1 Hypothesis testing with WMMD

As we have mentioned, from the statistical testing point of view, the coincidence axiom of the WMMD is key, as it ensures consistency against any alternative hypothesis $\mathcal{H}_1$. Then, given a significance level α for the two-sample test, a test can be constructed such that $\mathcal{H}_0$ is rejected when $\widehat{\text{WMMD}}^2 > r$.

The expected behaviour of $\widehat{\text{WMMD}}^2$ under the null which we might use to define r however differs from conventional bounds used for U-statistics. The reason is that in practice weights are data-dependent and have their own asymptotic behaviour which needs to be accounted for. In this case, under mild conditions that ensure well defined limits for these weights, also the asymptotic distributions are well defined. This result is given in Theorem 1 below.

Theorem 1 (Asymptotic distribution of WMMD). *Assume that k has finite second moments and that the weight matrix $W \in \mathbb{R}^{n\times n}$ ($W_{ij} = w(x_i)w(x_j)$) be approximately diagonalizable (made precise in Appendix B.1). Then, the following statements hold,*

1. *Under $\mathcal{H}_0$, the asymptotic distribution of $\widehat{WMMD}^2$ is given by a mixture of independent χ^2 random variables. We provide the exact terms in Appendix B.1.*
2. *Under $\mathcal{H}_1$,*

$$n^{1/2}\left(\widehat{WMMD}^2 - WMMD^2\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\mathcal{H}_1}\right).$$

We have used $\xrightarrow{d}$ to denote convergence in distribution. See Appendix B.1 for concrete expressions of all terms involved and a proof that relies on an approximate eigen-decomposition of the weight matrix and involves large-sample distributional approximations of quadratic forms and $U-$statistics.

3.2.2 Approximating the weights in practice

While we have shown that our test statistic is consistent against all alternatives, in practice simulating from the asymptotic null distribution can be challenging. The distribution under

the null requires knowledge of the sample selection mechanism, that is the design densities of the assignment variable T in the two populations, which is not available. A straightforward solution is to estimate each function $Pr(T=1|X=x)$ and $Pr(T=0|X=x)$ separately, for example with a classification algorithm, although this has been shown to result in unstable estimates of the ratio $Pr(T=1|X=x)/Pr(T=0|X=x)$ when the denominator is small [168] and adds an additional computational burden to the test procedure. An alternative approach is to use a plug-in estimate for the ratio directly. The approach we take is to estimate weights $\hat{w}(x)$ such that $Pr(T=1|X=x) \approx \hat{w}(x)Pr(T=0|X=x)$ by matching feature representation of both domains in a high-dimensional feature space [74].

We estimate weights $\hat{w}$ such as to minimize the distance between mean embeddings in a RKHS $\mathcal{H}_K$ with kernel K that is defined implicitly by a feature map $\phi : \mathcal{X} \to \mathcal{H}_K$ of the confounding variable distributions in the two populations,

$$\hat{w} := \underset{0<w<B}{\operatorname{argmin}} \left|\left| \mathbb{E}_{P_{X|T=0}} w(x)\phi(x) - E_{P_{X|T=1}} \phi(x) \right|\right|_{\mathcal{H}_K}. \tag{3.6}$$

This problem is convex. For injective mappings, minimizing (3.6) converges to $Pr(T=1|X=x)/Pr(T=0|X=x)$ and $\hat{w}$ can be found with a quadratic program for which many efficient solvers have been developed. In our implementation we use the Gaussian kernel with bandwidth parameter set to the median Euclidian distance between values of the confounding variables. Theorem 2 below guarantees that the density ratio estimation using (3.6) in the computation of $\widehat{\text{WMMD}}$ and of the asymptotic null distribution still yields a consistent test.

Theorem 2 (Consistency of $\widehat{\text{WMMD}}$). *Let $\hat{w}(x)$ be the empirical density ratio estimates of $w(x)$ – the underlying population value – derived by matching the kernel mean embeddings of the observed distributions of confounding variables $P_{X|T=1}$ and $P_{X|T=0}$. Suppose the test threshold is set to the upper α quantile of the distribution of the WMMD under $\mathcal{H}_0$. Then, asymptotically, the false positive rate with estimated weights is α and its power converges to 1.*

The proof, given in Appendix B.1, is based on the consistency of kernel mean matching to approximate the likelihood ratio in the asymptotic regime. While importance weighting using the likelihood ratio results in $\widehat{\text{WMMD}}$ being an asymptotically unbiased estimator of the MMD, the estimator may not concentrate well because the weights may be large or inaccurate due to the finite samples available in practice. We now provide a concentration bound for $\widehat{\text{WMMD}}$ for the case where weights are upper-bounded by some maximum value.

Theorem 3 (Large deviation bound of $\widehat{\text{WMMD}}$). *Let $\{y_i, t_i, x_i\}_{i=1}^{n+m}$ be i.i.d observations drawn from the joint distribution of random variables (Y,T,X), n of them with $t_i = 1$ and m with $t_i = 0$.*

Assume the feature representation $\phi(x) \in H_K$ *be bounded above by R,* $w(x) \leq B$ *for all* $x \in \mathcal{X}$*, and that there exists an* $\varepsilon > 0$ *such that,*

$$\left\| \frac{1}{n} \sum_{i=1:t_i=1}^{n} \hat{w}(x_i)\phi(x_i) - \frac{1}{m} \sum_{i=1:t_i=0}^{m} \phi(x_i) \right\|_{H_K} \leq \varepsilon.$$

Then, with probability at least $1-\delta$*, the absolute difference in estimation of weighted estimator* $\widehat{WMMD}$ *in comparison to the MMD,* $|\widehat{WMMD}^2 - MMD^2|$ *is bounded above by,*

$$2R(B+1)\left(\varepsilon + \left(1+\sqrt{2\log\frac{2}{\delta}}\right)R\sqrt{\frac{B^2}{n}+\frac{1}{m}}\right) + R(B+1)^2\sqrt{\frac{1}{2m_2}\log\frac{1}{\delta}},$$

where $m_2 := \lfloor m/2 \rfloor$.

Qualitatively, B measures the maximum allowed discrepancy between $Pr(T=1|X=x)$ and $Pr(T=0|X=x)$ (and is a user defined parameter in practice, we set it to 10 as a default in our experiments). A low value of B ensures robustness of the learned representations by limiting the influence of individual observations, thus reducing the variance of the resulting estimator and improving its concentration around the true estimate. However, with strong bias - the discrepancy between $Pr(T=1|X=x)$ and $Pr(T=0|X=x)$ is large - limiting B will result in higher ε which increases the bound. In turn, as expected, concentration improves with sample size. Asymptotically in m and n with high probability, the concentration of the representation depends only on matching confounding distributions in feature space $\mathcal{H}_K$. This shows that unbiased two-sample testing is not possible unless enough *comparable* examples in the two populations exist.

3.2.3 Connections with testing in regression models

There is a close connection between testing for distributional differences in two outcome samples independent of confounding and the predictive power of those factors on the outcome. In fact, adjustment is needed precisely because confounding variables are both predictive of the outcome and predictive of the sample selection mechanism. In one approach, the source of variation due to sample selection bias on the outcome y can be modelled explicitly, for example by considering a regression model with random effects. Consider the following random effect regression model [151] for the outcome y,

$$Y_i = \mu + Z_i u_i + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{3.7}$$

where $Z_i \in \{0,1\}$ represents the assignment of example i into one of the two samples and $u_i \sim \mathcal{N}(0,\sigma_u^2)$. Under the null assumption, testing for variation in Y that is irrelevant of the sample selection mechanism (which is our goal) is then equivalent to testing the variance component $\sigma_u^2 = 0$ [66, 107]. A score test statistic for this problem is given by $S = \sum_{i=1}^n \sum_{j=1,j\neq i}^n k_{ij}\tilde{Y}_i\tilde{Y}_j + \sum_{i=1}^n \tilde{Y}_i^2$ where $\tilde{Y}_i := \frac{(Y_i-\mu)}{\sigma}$, see e.g Section 4 in [66]. The statistic S therefore has a high value whenever the terms of the matrix $K = (k_{ij})$ and the matrix $\tilde{Y}\tilde{Y}^T = (\tilde{Y}_i\tilde{Y}_j)$ are correlated. Now consider the case $n = m$ and write $y_{i,1} = y_i$ such that $t_i = 1$, and analogously for $y_{j,0}$, $i, j = 1,\ldots,n$. Let k_{ij} be a column vector with entries $\left[k(y_{i,1},y_{j,1}), k(y_{i,0},y_{j,0}), k(y_{i,1},y_{j,0}), k(y_{i,0},y_{j,1})\right]$ and let w_{ij} have entries $\left[w(x_i)w(x_j), 1, -w(x_i), -w(x_j)\right]$. Then we may write,

$$\widehat{\mathrm{WMMD}}^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} w_{ij}^T k_{ij},$$

which can be interpreted as a non-linear alternative to the first term of S where the inner product $\langle a,b\rangle = a^T b$ is replaced by the inner product in feature space $k(a,b)$.

3.3 Related work

Many empirical studies, especially those investigating treatments and effects from finite samples *require* a notion of statistical significance to assess treatment outcomes. Classical tests for this problem are mostly local in nature i.e., testing for significance of estimated parameters in a regression model or concerned with average effects or average effects within defined subgroups and not with differences in the outcome distribution as a whole as considered in this chapter. Some examples include the weighted two-sample t-test (see e.g. [25]), the randomization test of [49] and the nonparametric tests for treatment effect heterogeneity of [43]. Extensions of other classical two-sample tests are also possible. One possible approach is using ANCOVA (Analysis of Covariance) methods which proceed by regressing the outcome variable on confounding variables before comparing the variation of the corresponding residuals between the two populations to the variation of the residuals within each one of the two populations, for example with an F-test [172]. These have the advantage of being more powerful in the settings where they apply but also restrict the class of alternative hypotheses [103].

With respect to two-sample tests for differences in any moment of the distribution existing tests, in some cases, may be adjusted to accommodate for selection bias. One extension to (full distribution) two-sample testing that may be considered for this problem is to first, partition the combined population into homogeneous subgroups (such that the feature distribution of

confounding variables approximately agree in each subgroup, for example using the propensity score) and second, compute two sample tests statistics in each subgroup before averaging their results. Such tests would take the form of block tests or B-tests [194], proposed initially as more efficient alternatives to conventional tests. In our experiments, we implement non-parametric versions of some of these tests and refer the reader to Appendix B.3 for a more detailed description of implementations in each case.

Outside the hypothesis testing literature, weighted statistics are frequent, often referred to as importance sampling techniques and inverse probability weighting methods [168]. Using importance weights with the MMD specifically has been used in generative models to sample from modified distributions [48] and for unsupervised domain adaptation [190, 191].

3.4 Experiments

In this section we compare two-sample tests on both artificial benchmark data and real-world data. The focus of our results will be on the evaluation of **power**: the rate at which we correctly reject $\mathcal{H}_0$ when it is false; and **type I error**: the rate at which we incorrectly reject $\mathcal{H}_0$ when it is true. $\alpha = 0.05$ throughout.

Baseline Tests. The proposed test is denoted **WMMD**. Comparisons are made with three tests. The **ANCOVA** F-test based on regression residuals from a random forest model. The block-based approach where partitions are made based on the propensity score and two-sample tests in each partition conducted with the MMD [194] (**Block-MMD**). The Block-MMD can be seen as an alternative adjusting for selection bias in subsets of the data separately, rather than continuously as with our approach and which we expect to have uncontrolled type I error in heterogeneous data samples. And finally, the unweighted (conventional) **MMD** test [70] that serves to measure the benefit of adjustments for selection bias as well as any loss in performance by using the WMMD in data that is not biased.

For kernel-based tests, since their null distributions are given by an infinite sum of weighted chi-squared variables (no closed-form quantiles), in each trial we use 400 random permutations to approximate the null. Details of implementations are given in the Appendix.

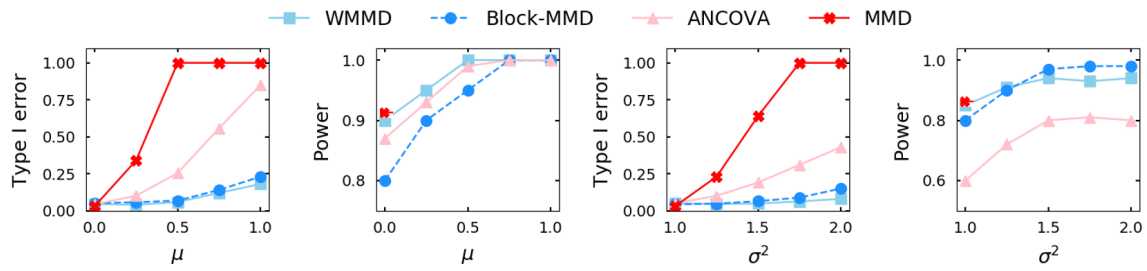


Figure 3.2: Type I error (lower better) and Power (higher better) of all tests on synthetic experiments. The WMMD has simultaneously best control of type I errors and highest power.

3.4.1 Synthetic examples

The primary objective of our synthetic simulations will be to analyse the influence of the sampling selection mechanism on performance. Here it will be particularly interesting to understand our test's behaviour on samples that appear different (in distribution) but only because of an underlying mismatch in confounding variables that simultaneously influence the distributions of interest. In this case we would expect conventional two sample tests to reject the null hypothesis resulting in uncontrolled type I error ($> \alpha$). And similarly for the case of observed distributions that seem to match (in distribution) due to spurious correlations that we show results in low power of traditional tests.

Experiment design. We consider the following data distributions for two samples of data ($T = 0$ and $T = 1$) that exhibit a spurious dependence between their respective outcome distributions $Y|T = 0$ and $Y|T = 1$ such as might occur due to selection bias,

$$X|T = 0 \sim \mathcal{N}(0, I), \qquad X|T = 1 \sim \mathcal{N}(\mu, \sigma^2 I),$$
$$Y|T = i \sim g_i(X) + \mathcal{N}(0, I), \quad i = 0, 1.$$

With this data generating mechanism, units in our two samples ($T = 0$ and $T = 1$) have differing confounder distributions $X|T$, a systematic difference which creates a spurious connection between T and Y.

Recall that the hypothesis testing problem is to evaluate, with data sampled from the model above, the evidence for a difference in distribution P_{Y^1} and P_{Y^0},

$$\mathcal{H}_0 : P_{Y^1} = P_{Y^0} \quad \text{versus} \quad \mathcal{H}_1 : P_{Y^1} \neq P_{Y^0}, \tag{3.8}$$

μ and σ^2 determine selection bias, i.e. the extent of the dependence between X and T which biases the dependence between T and Y. The distributions we are interested in discriminating are P_{Y^0} and P_{Y^1} (which reduces to $g_0 = g_1$ under the null, and $g_0 \neq g_1$ under the alternative), which implicitly remove selection bias by breaking the dependency between X and T.

Performance with increasing bias

In a first experiment we investigate the influence of increasing selection bias with two problems:

1. Difference in means μ (with $\sigma^2 = 1$) of confounding variables across the two samples. Results in the two left-most panels of Figure 4.2.
2. Difference in variances σ^2 (with $\mu = 0$) of confounding variables across the two samples. Results in the two right-most panels of Figure 4.2.

In each case the dimensionality of X and Y are set to 20, the number of samples in each population to $n = 400$. Under $\mathcal{H}_0$, $g_0(x) = g_1(x) = x + x^2$, and under $\mathcal{H}_1$, $g_0(x) = x$ and $g_1(x) = [\sin(x_1), x_2, ..., x_{20}]$. This set-up is designed to be a challenging problem with moderately high-dimensionality, non-linear dependencies and for the alternative hypothesis differences only in the first dimension of X.

Results. Across experiments (Figure 4.2) WMMD is the only test that successfully adjusts for selection bias, with controlled type I error even in relatively high bias settings (for instance for $\mu = 1$, only 60% of their densities overlap) while other alternatives underperform.

As anticipated, conventional two-sample test such as the MMD fail with the presence of confounders, we omit plotting the MMD for the power results (beyond $\mu = 0$ and $\sigma^2 = 1$) due to its poor type I error control. We notice also that the Type I error of the block-MMD deteriorates substantially for the variance experiment, potentially because a coarse partition may introduce artificial differences between samples that lead the test to reject the null more often than desired. The panels describing power show good performance for all methods. It is also expected that power increases with confounder distributional shift, as it results in more divergent outcome distributions (and thus easier to distinguish). However, unless type I error is controlled, those results lose their significance. Among methods that control type I error (WMMD and Block-MMD for low bias settings i.e. first half of each panel approximately), WMMD has higher or competitive power.

We make an important comparison also in the two power experiments in the absence of selection bias (the point where the MMD in red is computed). The MMD and WMMD have comparable

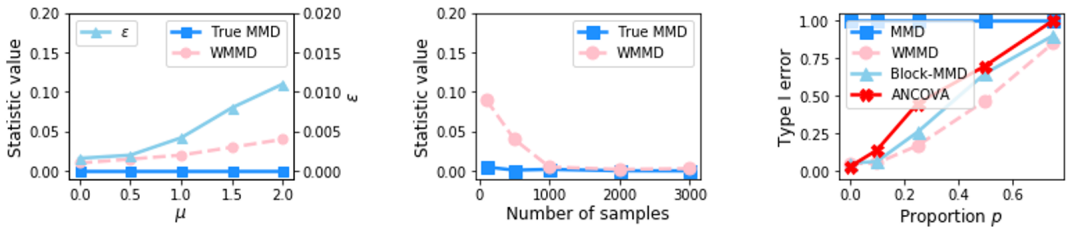


Figure 3.3: The two leftmost panels show the approximation error of the WMMD with increasing bias and increasing sample size. The rightmost panel show type I errors in the presence of unobserved confounders.

performance, which suggests that the WMMD is *almost as efficient* as the MMD in datasets tailored to the latter (when no bias exists), while also having good performance in the presence of bias. This is important because in most cases it is not known which variables confound the association between group membership and outcome. What this result means is that we are not worse-off using the WMMD even when there is no selection bias. In this sense the WMMD generalizes the MMD.

Relating to our theoretical results

Even though performing competitively, we observe the WMMD to loosen control of type I error as the strength of bias increases. In the following experiments we consider data generated under $\mathcal{H}_0$ as described in the first paragraph of section 3.4.1. and investigate the estimated WMMD statistic in comparison with optimal behaviour (defined as "True MMD", that is the MMD computed from data with no unobserved confounding on distributions Y^0, Y^1 not accessible in practice).

Results. With increasing confounding, we see in the leftmost panel of Figure 3.3 that the WMMD departs from its optimal value. The reason is that matching distributions of confounders gets harder with increasing confounding. Notice for instance the increasing value of ε in the opposite vertical axis, that quantifies the difference between matched distributions introduced in Theorem 3. The middle panel shows however that this discrepancy rapidly vanishes with increasing sample size. Here, we have fixed $\mu = 1$ and increased the sample size to see the estimation error converging to zero.

The takeaway is that a larger number of samples can be expected to be required to successfully control for type I errors to the desired threshold, while the number of samples depends on the strength of the confounding bias among the two samples.

What if confounding is unobserved?

We have assumed until now that the selection bias is completely driven by factors available to the researcher. In most real applications this will not be the case. We simulate such a scenario by including unobserved confounders in the sample selection mechanism under the null with the same specifications considered above. To do so, we hide or remove from the observed data a proportion p of variables X.

Results. The results, as a function of p, are shown in the rightmost panel of Figure 3.3. Unobserved confounders introduce variation in the outcome distribution that cannot be adjusted for since it is unobserved, which translates in uncontrolled type I errors for all methods. One may not expect to consistent hypothesis testing in this scenario but we note that this criticism extends to all methods with an assumption of ignorability, and in particular including most treatment effect estimation algorithms.

Remark. Variables X, treated as confounders in our case, may play other roles in general graphical models, for example as mediators or colliders (in both cases with an arrow from T into X). For the purposes of two-sample testing of treatment effects however we may rule out both of these cases because of temporal precedence, i.e. we cannot have an arrow going from T into X because group (treatment) assignment is done *after* observation of X. In others, if T represents a pre-existing characteristic of individuals (such as gender in the in the example of the introduction) we must validate the causal graph to ensure correct conclusions.

3.4.2 Employment program evaluation

The problem is to determine the effectiveness of an employment program implemented in the mid-1970s in the U.S. to individuals who had faced economic and social hardship [99]. The outcome of interest is earnings two years after the end of the employment program. Our null hypothesis is no difference in earnings with the program, with respect to earnings without the program. Posterior earning in treated and control populations are not directly comparable because the populations differ systematically in their education level, prior earnings, age, ethnicity and marital status: all plausible confounders. The data contains 614 individuals, 185 of whom were included in the employment program.

Table 3.1: Type I error at level $\alpha = 0.05$ as a function of artificially introduced bias p.

p	**0.05**	**0.10**	**0.15**	**0.20**
MMD	0.95	1	1	1
Block-MMD	0.051	0.055	0.070	0.083
ANCOVA	0.045	0.040	0.056	0.096
WMMD	0.051	0.043	0.052	0.060

With real data, the ground truth relationship between two populations is unknown. To compare the performance of our test, however, we can simulate a distribution under the null $\mathcal{H}_0$ by shuffling all variables into two populations, and subsequently introducing bias by selectively removing observations based on a set of confounding covariates. To remove observations, we build a linear regression model to predict earnings based on confounding variables and remove those observations with *high* predicted earnings in one group and those with *low* predicted earning in the other group. After adjusting for this bias the two populations should be equal in distribution and performance comparisons are then made in terms of type I error. A similar approach is used for conventional two sample testing, see for example the experiments in [111]. Type I error as a proportion p of observations removed (that is increasing bias) is given in Table 3.1. On the original data, all tests returned significant earning difference. This is an important result in its own right as it demonstrates an effect independent of confounding bias.

3.5 Conclusions

We have proposed a test statistic for the two-sample problem that expands the toolkit of statisticians to make inference on treatment effects with selection-biased data. Bias in the sample selection mechanism creates distributional shift which leads to bias in the treatment effect if unaccounted for. Making inference on the significance of treatment effects in this context is challenging and under-explored. To our knowledge, our test is the first to consider two-sample testing in biased groups of data.

Our proposal is a generalization of the MMD to adjust for this bias. We have demonstrated our test to be consistent in the presence of selection bias, derived its asymptotic distribution and derived large deviation bounds of approximations in practice. In empirical comparisons, we have shown our test to be more powerful than existing alternatives while controlling approximately for type I error.

The weighting strategy and proof techniques presented in this chapter are not specific to the two sample problem and may be applied to kernel-based tests for other problems, such as independence testing [69], conditional independence testing [197] and three variable interaction testing [154]. Similarly, one may extend the proposed approach to test and adjust for selection bias in other structured spaces where kernels are known to be characteristic such as other compact metric spaces [12].

Chapter 4

Conditional Independence Testing using Generative Adversarial Networks

In this chapter, we consider the hypothesis testing problem of detecting conditional dependence, with a focus on high-dimensional feature spaces. Our contribution is a new test statistic based on samples from a generative adversarial network designed to approximate directly a conditional distribution that encodes the null hypothesis, in a manner that maximizes power (the rate of true negatives). We show that such an approach requires only that density approximation be viable in order to ensure that we control type I error (the rate of false positives); in particular, no assumptions need to be made on the form of the distributions or feature dependencies. Using synthetic simulations with high-dimensional data we demonstrate significant gains in power over competing methods. In addition, we illustrate the use of our test to discover causal markers of disease in genetic data.

Introduction

Conditional independence tests are concerned with the question of whether two variables X and Y behave independently of each other, after accounting for the effect of confounders Z. Such questions can be written as a hypothesis testing problem:

$$\mathcal{H}_0 : X \perp\!\!\!\perp Y | Z \quad \text{versus} \quad \mathcal{H}_1 : X \not\perp\!\!\!\perp Y | Z.$$

Tests for this problem have recently become increasingly popular in the Machine Learning literature [156, 197, 155, 147, 50] and find natural applications in causal discovery studies in all areas of science [100, 131]. An area of research where such tests are important is genetics,

where one problem is to find genomic mutations directly linked to disease for the design of personalized therapies [203, 93]. In this case, researchers have a limited number of data samples to test relationships even though they expect complex dependencies between variables and often high-dimensional confounding variables Z. In settings like this, existing tests may be ineffective because the accumulation of spurious correlations from a large number of variables makes it difficult to discriminate between the hypotheses. As an example the work in [138] shows empirically that kernel-based tests have rapidly decreasing power with increasing data dimensionality.

In this chapter, we present a test for conditional independence that relies on a different set of assumptions that we show to be more robust for testing in high-dimensional samples (X,Y,Z). In particular, we show that given only a viable approximation to a conditional distribution one can derive conditional independence tests that are approximately valid in finite samples and that have non-trivial power. Our test is based on a modification of Generative Adversarial Networks (GANs) [67] that simulates from a distribution under the assumption of conditional independence, while maintaining good power in high dimensional data. In our procedure, after training, the first step involves simulating from our network to generate data sets consistent with $\mathcal{H}_0$. We then define a test statistic to capture the $X-Y$ dependency in each sample and compute an empirical distribution which approximates the behaviour of the statistic under $\mathcal{H}_0$ and can be directly compared to the statistic observed on the real data to make a decision.

The chapter is outlined as follows. In section 4.1, we provide an overview of conditional hypothesis testing and related work. In section 4.2, we provide details of our test and give our main theoretical results. Sections 4.3 and 4.4 provide experiments on synthetic and real data respectively, before concluding in section 4.5.

4.1 Background

We start by introducing our notation and review central notions of hypothesis testing. Throughout, we will assume the observed data consists of n $i.i.d$ tuples (X_i,Y_i,Z_i), defined in a potentially high-dimensional space $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$, typically $\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}\times\mathbb{R}^{d_z}$. Conditional independence tests statistics $T:\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}\to\mathbb{R}$ summarize the evidence in the observational data against the hypothesis $\mathcal{H}_0: X \perp\!\!\!\perp Y|Z$ in a real-valued scalar. Its value from observed data, compared to a defined threshold then determines a decision of whether to reject the null hypothesis $\mathcal{H}_0$ or not reject $\mathcal{H}_0$. Hypothesis tests can fail in two ways:

- Type I error: rejecting $\mathcal{H}_0$ when it is true.
- Type II error: not rejecting $\mathcal{H}_0$ when it is false.

We define the p-value of a test as the probability of making a type I error, and its power as the probability of correctly rejecting H_0 (that is 1 - Type II error). A good test requires the p-value to be upper-bounded by a user defined significance level α (typically $\alpha = 0.05$) and seeks maximum power. Testing for conditional independence is a challenging problem. Shah et al. [159] showed that no conditional independence test maintains non-trivial power while controlling type I error over any null distribution. In high dimensional samples (relative to sample size), the problem of maintaining good power is exacerbated by spurious correlations which tend to make X and Y appear independent (conditional on Z) when they are not.

4.1.1 Related work

A recent favoured line of research has characterized conditional independence in a **reproducing kernel Hilbert space** (RKHS) [197, 50]. The dependence between variables is assessed considering all moments of the joint distributions which potentially captures finer differences between them. [197] uses a measure of partial association in a RKHS to define the KCIT test with provable control on type I error asymptotically in the number of samples. Numerous extensions have also been proposed to remedy high computational costs, such as [167] that approximates the KCIT with random Fourier features making it significantly faster. Computing the limiting distribution of the test becomes harder to accurately estimate in practice [197], and different bandwidth parameters give widely divergent results with dimensionality [138], which affects power.

To avoid tests that rely on asymptotic null distributions, **sampling strategies** consider explicitly estimating the data distribution under the null assumption $\mathcal{H}_0$. Permutation-based methods [50, 147, 20, 156] follow this approach. To induce conditional independence, they select permutations of the data that preserve the marginal structure between X and Z, and between Y and Z. For a set of continuous conditioning variables and for sizes of the conditioning set above a few variables, the "similar" examples (in Z) that they seek to permute are hard to define as common notions of distance increase exponentially in magnitude with the number of variables. The approximated permutation will be inaccurate and its computational complexity will not be manageable for use in practical scenarios. As an example, [50] constructs a permutation P that enforces invariance in Z ($PZ \approx Z$) while [147] uses nearest neighbors to define suitable permutation sets.

We propose a different sampling strategy building on the ideas proposed by [30] that introduce the conditional randomization test (CRT). It assumes that the conditional distribution of X given Z is known under the null hypothesis (in our experiments we will assume it to be Gaussian for use in practice). The CRT then compares the known conditional distribution to the distribution of the observed samples of the original data using summary statistics. Instead we require a weaker assumption, namely having access to a viable approximation, and give an approximately valid test that does not depend on the dimensionality of the data or the distribution of the response Y; resulting in a non-parametric alternative to the CRT. [20] also expands the CRT by proposing a permutation-based approach to density estimation.

Generative adversarial networks have been used for hypothesis testing in [155]. In this work, the authors use GANs to model the data distribution and fit a classification model to discriminate between the true and estimated samples. The difference with our test is that they provide only a loose characterization of their test statistic's distribution under $\mathcal{H}_0$ using Hoeffding's inequality. As an example of how this might impact performance is that Hoeffding's inequality does not account for the variance in the data sample which biases the resulting test. A second contrast with our work is that we avoid estimating the distribution exactly but rather use the generating mechanism directly to inform our test.

4.2 Generative Conditional Independence Test

Our test for conditional independence, the GCIT (short for Generative Conditional Independence Test), compares an observed sample with a generated sample equal in distribution if and only if the null hypothesis holds. We use the following representation under $\mathcal{H}_0$,

$$Pr(X|Z,Y) = Pr(X|Z) \sim q_{\mathcal{H}_0}(X). \tag{4.1}$$

On the right hand side the null model preserves the dependence structure of $Pr(X,Z)$ but breaks any dependency between X and Y. If actually there exists a direct causal link between X and Y then replacing X with a null sample $\tilde{X} \sim q_{\mathcal{H}_0}$ is likely to break this relationship.

Sampling repeatedly $\tilde{X}$ conditioned on the observed confounders Z results in an exchangeable sequence of generated triples $(\tilde{X},Y,Z)$ and original data (X,Y,Z) under $\mathcal{H}_0$. In this context, any function ρ – such as a statistic $\rho : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ – chosen independently of the values

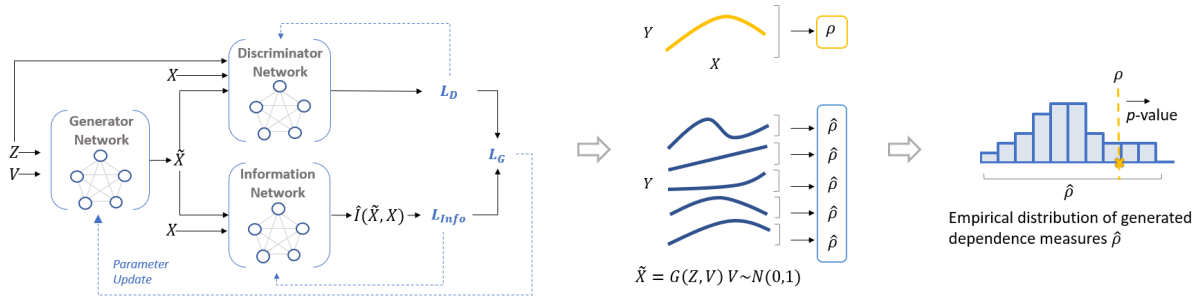


Figure 4.1: Illustration of conditional independence testing with the GCIT. A generator G is optimized by adversarial training to estimate the conditional distribution $X|Z$ under $\mathcal{H}_0$. We then use G to generate synthetic samples of $\tilde{X}$ under the estimated conditional distribution. Multiple draws are taken for each configuration Z and a measure of dependence between generated $\tilde{X}$ and Y, $\hat{\rho}$, is computed. The sequence of synthetic $\hat{\rho}$ is subsequently compared to the original sample statistic ρ to get a p-value.

of X applied to the real and generated samples preserves exchangeability. Hence the sequence,

$$\rho(X,Y,Z), \rho(\tilde{X}^{(1)},Y,Z), \ldots, \rho(\tilde{X}^{(M)},Y,Z), \tag{4.2}$$

is exchangeable under the null hypothesis $\mathcal{H}_0$, deriving from the fact that the observed data is equally likely to have arisen from any of the above. Without loss of generality, we assume that larger values of ρ are more extreme. The p-value of the test can be approximated by comparing the generated samples with the observed sample,

$$\sum_{m=1}^{M} \frac{1 + \mathbf{1}\{\rho(\tilde{X}^{(m)},Y,Z) \geq \rho(X,Y,Z)\}}{M+1}, \tag{4.3}$$

which can be made arbitrarily close to the true probability, $\mathbb{E}_{\tilde{X}\sim q_{\mathcal{H}_0}} \mathbf{1}\{\rho(\tilde{X},Y,Z) \geq \rho(X,Y,Z)\}$, by sampling additional features $\tilde{X}$ from $q_{\mathcal{H}_0}$. $\mathbf{1}$ is the indicator function. Figure 4.1 gives a graphical overview of the GCIT.

4.2.1 Generating samples from $q_{\mathcal{H}_0}$

In this section we describe a sampling algorithm that adapts generative adversarial networks [67] to generate samples $\tilde{X}$ conditional on high dimensional confounding variables Z. GANs provide a powerful method for general-purpose generative modeling of datasets by designing a discriminator D explicitly used as an adversary to train a generator G responsible for estimating $q_{\mathcal{H}_0} := Pr(X|Z)$. Over successive iterations both functions improve based on the performance of the adversarial player.

Our implementation is based on Energy-based generative neural networks introduced in [200] which if trained optimally, can be shown to minimize a measure of divergence between probability measures that directly relates to a theoretical bound shown in this section that underlies our method. Pseudo-code for the GCIT and full details on the implementation are given in Appendix C.4.

Discriminator. We define the discriminator as a function $D_\eta : \mathcal{X} \times \mathcal{Z} \to [0,1]$ parameterized by η that judges whether a generated sample $\tilde{X}$ from G is likely to be distributed as its real counterpart X or not, conditional on Z. We train the discriminator by gradient descent to minimize the following loss function,

$$\mathcal{L}_D := \mathbb{E}_{x\sim q_{\mathcal{H}_0}} D_\eta(x,z) + \mathbb{E}_{\tilde{v}\sim p(v)} \left(1 - D_\eta(G_\phi(v,z),z\right), \tag{4.4}$$

where $G_\phi(z,v), v \sim p(v)$ is a synthetic sample from the generator (described below) and $x \sim q_{\mathcal{H}_0}$ is a sample from the data distribution under $\mathcal{H}_0$. Note that in contrast to [200] we set the image of D to lie in $(0,1)$ and include conditional data generation.

Generator. The generator, G, takes (realizations of) Z and a noise variable, V, as inputs and returns $\tilde{X}$, a sample from an estimated distribution $X|Z$. Formally, we define $G : \mathcal{Z} \times [0,1]^d \to \mathcal{X}$ to be a measurable function (specifically a neural network) parameterized by ϕ, and V to be d-dimensional noise variable (independent of all other variables). For the remainder of the chapter, let us denote $\tilde{x} \sim \hat{q}_{\mathcal{H}_0}$ the generated sample under the model distribution implicitly defined by $\hat{x} = G_\phi(v,z), v \sim p(v)$. In opposition to the discriminator, G is trained to minimize

$$\mathcal{L}_G(D) := \mathbb{E}_{\tilde{x}\sim\hat{q}_{\mathcal{H}_0}} D_\eta(\tilde{x},z) - \mathbb{E}_{x\sim q_{\mathcal{H}_0}} D_\eta(x,z). \tag{4.5}$$

We estimate the expectations empirically from real and generated samples.

4.2.2 Validity of the GCIT

The following result ensures that our sampling mechanism leads to a valid test for the null hypothesis of conditional independence.

Proposition 1 (Exchangeability) *Under the assumption that $X \perp\!\!\!\perp Y|Z$, any sequence of statistics $(\rho_i)_{i=1}^M$ functions of the generated triples $(\tilde{X}^{(m)}, Y, Z)_{m=1}^M$ is exchangeable.* □

Proof. All proofs are given in Appendix C.3.

Generating conditionally independent samples with a neural network preserves exchangeability of input samples and thus leads to a valid p-value, defined in eq. (4.3), for the hypothesis of conditional independence. Under the assumption that the conditional distribution $q_{\mathcal{H}_0}$ can be estimated exactly, this implies that we maintain an exact control of the type I error in finite samples. In practice however, limited amounts of data and noise will prevent us from learning the conditional distribution exactly.

In such circumstances we show below that the excess type I error - that is the proportion of false negatives reported above a specified tolerated level α - is bounded by the loss function $\mathcal{L}_G$; which, moreover, can be made arbitrarily close to 0 for a generator with sufficient capacity. We give this second result as a corollary of the GAN's convergence properties in Appendix C.3.

Theorem 1 *An optimal discriminator D^* minimizing $\mathcal{L}_D$ exists; and, for any statistic $\hat{\rho} = \rho(X,Y,Z)$, the excess type I error over a desired level α is bounded by $\mathcal{L}_G(D^*)$,*

$$Pr(\hat{\rho} > c_\alpha | \mathcal{H}_0) - \alpha \leq \mathcal{L}_G(D^*), \tag{4.6}$$

where $c_\alpha := \inf\{c \in \mathbb{R} : Pr(\hat{\rho} > c) \leq \alpha\}$ is the critical value on the test's distribution and $Pr(\hat{\rho} > c_\alpha | \mathcal{H}_0)$ is the probability of making a type I error. □

Theorem 1 shows that the GCIT has an increase in type I error dependent only on the quality of our conditional density approximation, given by the loss function with respect to the generator, even in the worst-case under *any* statistic ρ. For reasonable choices of ρ, robust to errors in the estimation of the conditional distribution, this bound is expected to be tighter. The *key* assumption to ensure control of the type I error, and therefore to ensure the validity of the GCIT, thus rests solely on our ability to find a viable approximation to the conditional distribution of $X|Z$. The capacity of deep neural networks and their success in estimating heterogeneous conditional distributions even in high-dimensional samples make this a reasonable assumption, and the GCIT applicable in a large number of scenarios previously unexplored.

4.2.3 Maximizing power

For a fixed sample size, conditional dependence $\mathcal{H}_1 : X \not\perp\!\!\!\perp Y | Z$, is increasingly difficult to detect with larger conditioning sets (Z) as spurious correlations due to sample size make X and Y appear independent. To maximize power it will be desirable that differences between generated samples $\tilde{X}$ (under the model $Pr(X|Z)$) and observed samples X (distributed according

to $Pr(X|Z,Y)$) be as apparent as possible. In order to achieve this we will encourage $\hat{X}$ and X to have low mutual information because irrespective of dimensionality, mutual information between distributions in the null and alternative relates directly to the hardness of hypothesis testing problems, which can be seen for example via Fano's inequality (section 2.11 in [175]). To do so, we investigate the use of the information network proposed in [10] and used in the context of feature selection in [89]. [10] propose a neural architecture and training procedure for estimating the mutual information between two random variables. We approximate the mutual information with a neural network $T_\theta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, parameterized by θ, with the following objective function (to be maximized),

$$\mathcal{L}_{Info} := \sup_{\theta} \mathbb{E}_{p^{(n)}_{x,\tilde{x}}}[T_\theta] - \log \mathbb{E}_{p^{(n)}_{x} \times p^{(n)}_{\tilde{x}}}[\exp(T_\theta)]. \tag{4.7}$$

We estimate T_θ in alternation with the discriminator and generator given samples from the generator in every iteration. We modify the loss function for the generator to include the mutual information and perform gradient descent to optimize the generator on the following objective,

$$\mathcal{L}_G(D) + \lambda \mathcal{L}_{Info}. \tag{4.8}$$

$\lambda > 0$ is a hyperparameter controlling the influence of the information network. This additional term $(\lambda \mathcal{L}_{Info})$ encourages the generation of samples $\tilde{X}$ as independent as possible from the observed variables X such that the resulting differences (between $\tilde{X}$ and X) are truly a consequence of the *direct* dependence between X and Y rather than spurious correlations with confounders Z.

To provide some further intuition, one can see why generating data different than the sample observed in the alternative $\mathcal{H}_1$ might be beneficial by considering the following bound (proven in Appendix C.3),

$$\text{Type I error} + \text{Type II error} \geq 1 - \delta_{TV}(\hat{q}_{\mathcal{H}_0}, q_{\mathcal{H}_1}), \tag{4.9}$$

where $\hat{q}_{\mathcal{H}_0}$ is the estimated null distribution with the GCIT, $q_{\mathcal{H}_1}$ is the distribution under $\mathcal{H}_1$ and where δ_{TV} is the total variation distance between probability measures. This result suggests that when emphasizing the differences between the estimated samples and true samples from $\mathcal{H}_1$, which increases the total variation, can improve the overall performance profile of our test by reducing a lower bound on type I and type II errors.

Remark. The GCIT aims at generating samples whose conditional distribution matches the distribution of its real counterparts, but can be independent otherwise. It is that gap that the

power maximizing procedure intends to exploit. In practice, there will be a trade-off between the objectives of the discriminator and information network but we found that setting $\lambda = 10$ in our experiments achieved good performance. It should be noted also that hyperparameter selection cannot be performed using cross-validation as we do not have access to ground truth and so the hyperparameters must typically be fixed a priori. However, we can consider artificially inducing conditional independence $(X \perp\!\!\!\perp Y|Z)$ (by permuting variables X and Y such as to preserve the marginal dependence in (X,Z) and (Y,Z)) and choose hyperparameters that best control for type I error. We explore this further in Appendix C.1 and test configurations of λ with synthetic data in section 4.3.2.

4.2.4 Choice of statistic ρ

The bound on the type I error given in Theorem 1 holds for any choice of statistic ρ as it depends solely on the conditional distribution estimation. For choices of ρ less sensitive to spurious differences between generated and true samples when the null $\mathcal{H}_0$ holds, the type I error is expected to be below this bound. We experimented with various dependence measures (between two samples) as choices for ρ. We consider the Maximum Mean Discrepancy [70], Pearson's correlation coefficient, the distance correlation (which measures both linear and nonlinear association, in contrast to Pearson's correlation), the Kolmogorov-Smirnov distance between two samples and the randomized dependence coefficient [112]. In our experiments we use the distance correlation and analyze performance using all other measures in Appendix C.1.

4.3 Synthetic data example

In this section we analyse the performance of the GCIT in a controlled fashion with synthetic data against a wide range of competing algorithms, illustrating the effects of different components of our method. We consider the CRT [30] with pre-specified Gaussian sampling distribution, whose parameters are estimated from data; the kernel-methods KCIT [197] and RCoT [167] with bandwith parameter estimated with the median of all pairwise distances between X and Y, a common choice in the literature; and the CCIT [156], which does not make prior assumptions on data distributions but was also not specifically designed for high-dimensional data.

When testing at level α, type I error should be as close as possible to α even though this might not be the case because of violated assumptions or approximations. An important consideration

in our discussion of power as we increase the dimensionality of Z, is the choice of alternatives $\mathcal{H}_1$. For instance, if the strength of the dependency between X and Y increases, the hypothesis testing problem will be made artificially easier and bias our conclusions with regards to data dimensionality, as observed also in [138]. In every synthetic experiment, we maintain the mutual information between X and Y approximately constant by first generating data and second estimating the mutual information before deciding to draw a new dataset, if the mutual information disagrees with the previous draw, or otherwise proceed with testing. We estimate the mutual information with a Gaussian approximation, $MI(X,Y) = -\frac{1}{2}\log(1-\hat{\rho}^2)$, where $\hat{\rho}$ is the linear correlation between X and Y.

4.3.1 Setup

We generate synthetic data according to the "post non-linear noise model" similarly to [197, 50, 167] that defines (X,Y,Z) under $\mathcal{H}_0$ and $\mathcal{H}_1$ as follows,

$$\mathcal{H}_0: \quad X = f(A_f Z + \varepsilon_f), \quad Y = g(A_g Z + \varepsilon_g), \tag{4.10}$$

$$\mathcal{H}_1: \quad Y = h(A_h Z + \alpha X + \varepsilon_h). \tag{4.11}$$

The matrix dimensions of $A_{(\cdot)}$ are such that X and Y are univariate, matrix entries as well as parameter α are generated at random in the interval $[0,1]$, and lastly, the noise variables $\varepsilon_{(\cdot)}$ are 0 on average with variance 0.025. The distributions of X, Y and ε, and the complexity of dependencies via f, g and g will be tuned carefully to make performance comparisons in three settings:

(1) Multivariate Gaussian
We set f, g and h to be the identity functions which induces linear dependencies, $Z \sim \mathcal{N}(0,\sigma^2)$, and $X \sim \mathcal{N}(0,\sigma^2)$ under $\mathcal{H}_1$ which results in jointly Gaussian data under the null and the alternative. Such a setting matches the assumptions of all methods and the interest of this study will be to provide a baseline for more complex scenarios.

(2) Multivariate Laplace
Kernel choice has a large impact on power, as we demonstrate in this setting. In this case, we set f, g and h as before but use a Laplace distribution to generate Z and X. The RBF kernel in this case overestimates the "smoothness" of the data. This study highlights the robustness of the GCIT in comparison to kernel-based methods which is important since hyperparameters cannot be tuned by cross-validation.

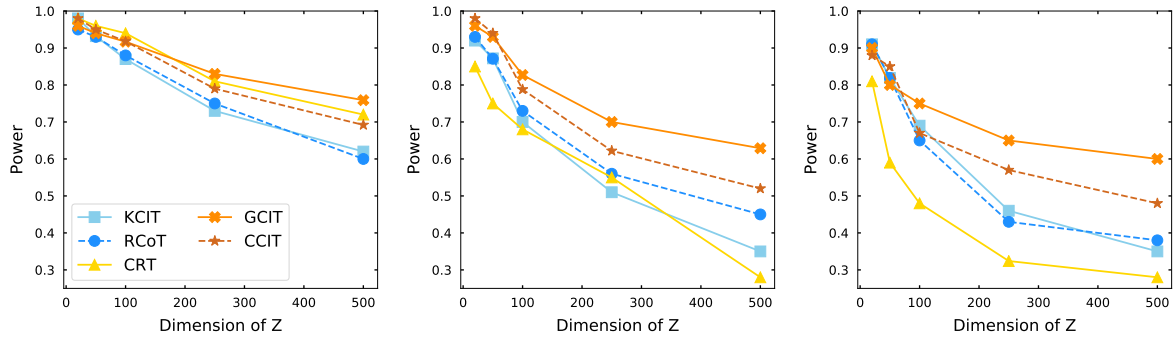


Figure 4.2: Power results of the synthetic simulations. (Higher better). **Left panel:** (1) Multivariate Gaussian, **Middle panel:** (2) Multivariate Laplace, **Right panel:** (3) Arbitrary distributions.

(3) Arbitrary distributions

We set f, g and h to be randomly sampled from $\{x^3, \tanh x, \exp(-x)\}$, resulting in more complex distributions and variable dependencies. Here $Z \sim \mathcal{N}(0, \sigma^2)$, and $X \sim \mathcal{N}(0, \sigma^2)$ under $\mathcal{H}_1$. This is our most general setting which most faithfully resembles the complexities we can expect in real applications.

Results: Power as a function of the dimensionality of Z is shown in Figure 4.2. Each point on the curves is computed by taking averages over 1000 random experiments with sample size equal to 500 examples. The results from scenario **(1)** are consistent with our expectations; all methods perform comparably, the CRT and kernel-based methods achieving high power in lower dimensions while slightly under-performing in higher dimensions. In scenario **(2)** and **(3)**, the failure of the CRT and kernel-based methods is apparent while the GCIT maintains high power, even with increasing dimensionality, which demonstrates the robustness of our sampling mechanism to arbitrary complex data distributions. The CCIT outperforms kernel-based methods in these cases also. An important contrast of the GCIT with respect to the CCIT is our addition of the information network, which we argue contributes to the higher power observed across all experiments. We analyze this empirically below.

Appendix C.2 shows that type I error is approximately controlled at a level α for all methods. Observe also that even though the GCIT requires training a new GAN in every iteration, in Appendix C.2 we show empirically that running times for the GCIT scale much better with dimensionality and sample size in comparison with the best benchmark, the CCIT: its running times are prohibitive in practice with more than 1000 samples or 500 dimensions in Z, with each test taking over $600s$ versus $60s$ for the GCIT.

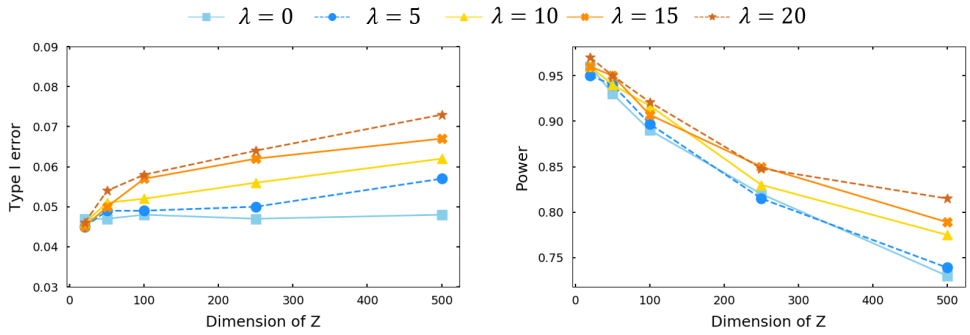


Figure 4.3: Type I error and power for different values of λ.

4.3.2 Source of gain: consequences of the information network

The information network aims to encourage maximum power in high-dimensional data. We control for its influence by varying λ in the loss function of the GCIT given in eq. (4.8). Higher values of λ encourage the generation of independent samples which improves power even though it might decrease the accuracy of the density approximation in the GAN optimization when the null in fact holds. We notice this trade-off between power and type I error for higher values of λ in Figure 4.3. The underlying data was generated from setting **(1)**, each curve in the two panels corresponds to a different value of λ. Lastly, we computed the lower-bound from GCIT generated samples and observed samples (by numerical integration) in eq. (4.9) to conclude that higher values of λ did decrease the lower bound, as expected.

4.4 Genetic data example

There is compelling evidence that the likelihood of a patient's cancer responding to treatment can be strongly influenced by alterations in the cancer genome [63]. We study the response of cancer cell lines to an anti-cancer drug where the problem is to distinguish between genetic mutations that influence directly the cancer cell line response from those that are not directly relevant [9, 173]. We use the subset of the CCLE data [9] relating to the drug PLX4720; it contains 474 cancer cell lines described by 466 genetic mutations. More details on the data can be found in Appendix C.5.

	AGREEMENT							DISAGREEMENT		
	BRAF.V600E	BRAF.MC	HIP1	CDC42BPA	THBS3	DNMT1	PRKD1	FLT3	PIP5K1A	MAP3K5
EN	1	3	4	7	8	9	10	5	19	78
RF	1	2	3	8	34	28	18	14	7	9
CRT	<0.001	<0.001	0.008	0.009	0.017	0.022	0.002	0.017	0.024	0.012
GCIT	<0.001	<0.001	0.008	0.050	0.013	0.020	0.002	0.521	0.001	<0.001

Figure 4.4: Genetic experiment results. Each cell gives the p-value or importance rank (where appropriate) indicating the dependency between a mutation and drug response.

Evaluating conditional independence relations from real data is difficult as we do not have access to the ground truth causal links. Instead we give our results in comparison to those of [9], who proceeded by reporting discriminative features returned by the parameter values of a fitted elastic net regression model (EN). This is common practice in genetic studies, see for example also [63]. In addition, we compare with the rank of each feature given by a random forest model importance scores (RF) and the p-value assigned by the CRT. The results for 10 selected mutations can be found in Figure 4.4. The first two rows give ranks of heuristic methods and the last two rows give p-values of conditional independence tests. We distinguish between the mutations where all methods agree (in the leftmost columns), and the mutations where not all methods agree (in the rightmost columns).

The mutations on genes PIP5K1A and MAP3K5 are recognized to be discriminative by the random forest model (high rank) and the GCIT (low p-value), which highlights the significance of the GCIT for conditional independence testing, suggesting that non-linear dependencies occur which are not captured by the elastic net or the CRT. For further evaluation, in this case we were able to cross-reference with a previous study to find evidence of the PIP5K1A gene to have a differential response on cancer cell lines when PLX4720 is applied [173]. The MAP3K5 gene has not previously been reported in the literature as being directly linked to the PLX4720 drug response, however [134] did find a proliferation of these gene mutations to be of BRAF *type* in cancer patients. This is interesting because PLX4720 is precisely designed as a BRAF inhibitor, and thus we would expect it to have an impact also on MAP3K5 mutations of the BRAF type. FLT3 is an interesting gene, found to be dependent on cancer response by the EN, RF and CRT, but not by the GCIT. This finding by the GCIT was confirmed however by a posterior genetic study [35] that established no link between cancer response and FLT3 mutations in the presence of PLX4720. Such results encourage us to believe that the GCIT is able to better detect dependence for these problems.

4.5 Conclusions

We proposed a generative approach to conditional independence testing using generative adversarial networks. We show this approach results in an approximately valid test for an arbitrary data distribution irrespective of the number of variables observed. We have demonstrated through simulated data significant gains in statistical power, and we illustrated the application of our method to discover genetic markers for cancer drug response on real high-dimensional data. From a practical perspective, algorithms based on other generative models can be constructed based on our proposed procedure that may be more adequate for different data modalities. In a general sense, this work opens the door to principled statistical testing with more heterogeneous data, and expands our ability to reason and test variable relationships in more challenging scenarios.

Part II

Causality with Heterogenous data

Chapter 5

Accounting for Unobserved Confounding in Domain Generalization

The ability to generalize from observed to new related environments is central to any form of reliable machine learning, yet most methods fail when moving beyond i.i.d data. This chapter argues that in some cases the reason lies in a misapreciation of the causal structure in data; and in particular due to the influence of unobserved confounders which void many of the invariances and principles of minimum error between environments presently used for the problem of domain generalization. This observation leads us to study generalization in the context of a broader class of interventions in an underlying causal model (including changes in observed, unobserved and target variable distributions) and to connect this causal intuition with an explicit distributionally robust optimization problem. From this analysis derives an new proposal for model learning with explicit generalization guarantees that is based on the partial equality of error derivatives with respect to model parameters. We demonstrate the empirical performance of our approach on healthcare data from different modalities, including image, speech and tabular data.

Introduction

Prediction algorithms use data, necessarily sampled under specific conditions, to learn correlations that extrapolate to new or related data. If successful, the performance gap between these two environments is small, and we say that algorithms *generalize* beyond their training data. Doing so is difficult however, some form of uncertainty about the distribution of new data is unavoidable. The set of potential distributional changes that we may encounter is mostly unknown and in many cases may be large and varied. Some examples include covariate shifts

[21], interventions in the underlying causal system [129], varying levels of noise [59] and confounding [127]. All of these feature in modern applications, and while learning systems are increasingly deployed in practice, generalization of predictions and their reliability in a broad sense remains an open question.

A common approach to formalize learning with uncertain data is, instead of optimizing for correlations in a *fixed* distribution, to do so simultaneously for a *range* of different distributions in an uncertainty set $\mathcal{P}$ [14].

$$\underset{f}{\text{minimize}} \sup_{P \in \mathcal{P}} \mathbb{E}_{(x,y)\sim P}[\mathcal{L}(f(x), y)], \tag{5.1}$$

for some measure of error $\mathcal{L}$ of the function f that relates input and output examples $(x, y) \sim P$. Choosing different sets $\mathcal{P}$ leads to estimators with different properties. It includes as special cases, for instance, many approaches in domain adaptation, covariate shift, robust statistics and optimization [97, 21, 53, 55, 162, 189, 1, 54]. Robust solutions to problem (5.1) are said to generalize if potential shifted, test distributions are contained in $\mathcal{P}$, but also larger sets $\mathcal{P}$ result in *conservative* solutions (i.e. with sub-optimal performance) on data sampled from distribution away from worst-case scenarios, in general.

One formulation of causality is in fact also a version of this problem, for $\mathcal{P}$ defined as any distribution arising from *arbitrary* interventions on observed covariates x leading to shifts in their distribution P_x (see e.g. sections 3.2 and 3.3 in [117]). The invariance to changes in covariate distributions of causal solutions is powerful for generalization, but implicitly assumes that all covariates or other drivers of the outcome subject to change at test time are observed. Often shifts occur elsewhere, for example in the distribution of unobserved confounders, in which case also conditional distributions $P_{y|x}$ may shift. Perhaps surprisingly, in the presence of unobserved confounders, the goals of achieving robustness and learning a causal model can be *different* (and similar behaviour also occurs with varying measurement noise).

There is, in general, an inherent *trade-off* in generalization performance. In the presence of unobserved confounders, causal and correlation-based solutions are both optimal in different regimes, depending on the shift in the underlying generating mechanism from which new data is generated. We consider next a simple example, illustrated in Figure 5.1, to show this explicitly.

Introductory example. We assume access to observations of variables (X_1, X_2, Y) in two training datasets, each dataset sampled with different variances ($\sigma^2 = 1$ and $\sigma^2 = 2$) from the

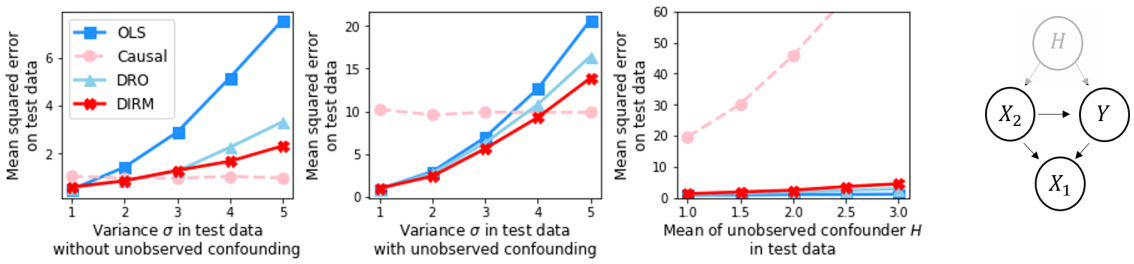


Figure 5.1: **The challenges of generalization**. Each panel plots testing performance under different shifts showing that there is an trade-off between causal and correlation-based solutions. The proposed approach, Derivative Invariant Risk Minimization (DIRM), is a relaxation of the causal solution that naturally interpolates between the causal solution and Ordinary Least Squares (OLS), and has, arguably, a more desirable risk profile.

following structural model $\mathbb{F}$,

$$X_2 := -H + E_{X_2}, \quad Y := X_2 + 3H + E_Y, \quad X_1 := Y + X_2 + E_{X_1},$$

$E_{X_1}, E_{X_2} \sim \mathcal{N}(0, \sigma^2)$, $E_Y \sim \mathcal{N}(0,1)$ are exogenous.

1. In a first scenario (**leftmost panel**) all data (training and testing) is generated *without* unobserved confounders, $H := 0$.

2. In a second scenario (**remaining panels**) all data (training and testing) is generated *with* unobserved confounders, $H := E_H \sim \mathcal{N}(0,1)$.

Each panel of Figure 5.1 shows performance on **new** data obtained after manipulating the underlying data generating system; the magnitude and type of intervention appears in the horizontal axis. We consider the following learning paradigms: Ordinary Least Squares (OLS) learns the linear mapping that minimizes average training risk, Domain Robust Optimization (DRO) minimizes the maximum training risk among the two available datasets, and the causal solution, assumed known, is the linear model with coefficients $(0,1)$ for (X_1, X_2). Two important observations motivate this chapter.

1. Observe that Ordinary Least Squares (OLS) and Domain Robust Optimization (DRO) absorb *spurious* correlations (due to H, and the fact that X_1 is caused by Y) with unstable performance under shifts in $p(X_1, X_2)$ but as a consequence good performance under shifts in $p(H)$. Causal solutions, by contrast, are robust to shifts in $p(X_1, X_2)$, even on new data with large shifts, but underperform substantially under changes in the distribution of unobserved confounders $p(H)$.

2. The presence of unobserved confounding *hurts* generalization performance in general with higher errors for all methods, e.g. contrast the y-axis of the middle and leftmost panel of Figure 5.1.

To the best of our knowledge, the influence of unobserved confounders has been minimally explored in the context of generalization of learning algorithms, even though, as Figure 5.1 shows, in this context different shifts in distribution may have important consequences for predictive performance.

Our Contributions. Our objective is to define a learning principle (5.1) that is robust to a wider set of distributions, accounting for the potential influence of unobserved confounders, and thus with a more desirable risk profile.

We start with the assumption that multiple training environments are available during training (e.g. patient data from multiple cohorts or hospitals). Using multiple environments, we show that certain statistical *invariances* across environments are expected in the presence of unobserved confounders, which motivate a new learning problem (5.1). Specifically, our proposal defines $\mathcal{P}$ in (5.1) as an *affine* combination of training data distributions and may be interpreted as an interpolation between purely correlation-based and causal-based solutions.

The practical benefits are threefold.

1. Solutions to this problem are provably robust to more general shifts in distribution, including shifts in observed, unobserved, and target variables, depending on the properties of the available training data distributions.
2. DIRM may be applied to any differentiable function.
3. We show significant out-of-sample performance gains and benefits for the reproducibility of variable selection with DIRM.

5.1 Invariances in the presence of unobserved confounders

This section formally introduces the problem of out-of-distribution generalization. We describe in greater detail the reasons that popular learning principles, such as Empirical Risk Minimization (ERM), underperform in general, and define certain invariances to recover solutions that generalize.

We take the perspective that all potential distributions that may be observed over a system of variables arise from a causal model $\mathcal{M} = (\mathbb{F}, \mathbb{V}, \mathbb{U})$, characterized by endogenous variables,

$\mathbb{V} \in \mathcal{V}$, representing all variables determined by the system, either observed or not; exogenous variables, $\mathbb{U} \in \mathcal{U}$, in contrast imposed upon the model, and a sequence of structural equations $\mathbb{F} : \mathcal{U} \to \mathcal{V}$, describing how endogenous variables can be (deterministically) obtained from the exogenous variables [129]. An example is given in Figure 5.1, $\mathbb{V} = (X_1, X_2, H, Y)$ are endogenous and $\mathbb{U} = (E_{X_1}, E_{X_2}, E_H, E_Y)$ are exogenous variables.

Unseen data is generated from such a system $\mathcal{M}$ after manipulating the distribution of exogenous variables $\mathbb{U}$, which propagates across the system shifting the joint distribution of all variables $\mathbb{V}$, whether observed or unobserved, but keeping the causal mechanisms $\mathbb{F}$ unchanged. Representative examples include changes in data collection conditions, such as due to different measurement devices, or new data sources, such as patients in different hospitals or countries, among many others.

Our goal is to learn a representation $Z = \phi(X)$ acting on a set observed variables $X \subset \mathbb{V}$ with the ability to extrapolate to new unseen data, and doing so acknowledging that all relevant variables in $\mathbb{V}$ are likely not observed. Unobserved confounders (for the task at hand, say predicting $Y \in \mathbb{V}$) simultaneously cause X and Y, confounding or biasing the causal association between X and Y giving rise to spurious correlations that do not reproduce in general [127, 129]. We present a brief argument below highlighting the systematic bias due to unobserved confounders in ERM.

5.1.1 The biases of unobserved confounding

Consider the following structural equation for observed variables (X, Y),

$$Y := f \circ \phi(X) + E, \tag{5.2}$$

where $f := f(\cdot; \beta_0)$ is a predictor acting on a representation $Z := \phi(X)$ and E stands for potential sources of mispecification and unexplained sources of variability. For a given sample of data (x, y) and $z = \phi(x)$, the optimal prediction rule $\hat{\beta}$ is often taken to minimize squared residuals, with $\hat{\beta}$ the solution to the normal equations: $\nabla_\beta f(z; \hat{\beta}) y = \nabla_\beta f(z; \hat{\beta}) f(z; \hat{\beta})$, where $\nabla_\beta f(z; \hat{\beta})$ denotes the column vector of gradients of f with respect to parameters β evaluated at $\hat{\beta}$. Consider the Taylor expansion of $f(z; \beta_0)$ around an estimate $\hat{\beta}$ sufficiently close to β_0, $f(z; \beta_0) \approx f(z; \hat{\beta}) + \nabla_\beta f(z; \hat{\beta})^T (\beta_0 - \hat{\beta})$. Using this approximation in our first order optimality condition we find,

$$\nabla_\beta f(z; \hat{\beta}) \nabla_\beta f(z; \hat{\beta})^T (\beta_0 - \hat{\beta}) + v = \nabla_\beta f(z; \hat{\beta}) \varepsilon, \tag{5.3}$$

where v is a scaled disturbance term that includes the rest of the linear approximation of f and is small asymptotically; $\varepsilon := y - f(z;\hat{\beta})$ is the residual. $\hat{\beta}$ is consistent for the true β_0 if and only if $\nabla_\beta f(z;\hat{\beta})\varepsilon \to 0$ in probability. This assumption is satisfied if E (all sources of variation in Y not captured by X) are independent of X (i.e. exogenous) or in other words if all common causes or confounders to both X and Y have been observed. Conventional regression may assign significant associations to variables that are neither directly nor indirectly related to the outcome, and in this case, we have no performance guarantees on new data with changes in the distribution of these variables. Omitted variables are a common source of unobserved confounding but we note in Appendix D.2 that similar biases also arise from other prevalent model mispecifications, such as measurement error [32].

5.1.2 Invariances with multiple environments

The underlying structural mechanism $\mathbb{F}$, that also relates unobserved with observed variables, even if unknown, is stable irrespective of manipulations in exogenous variables that may give rise to heterogeneous data sources. Under certain conditions, statistical footprints emerge from this structural invariance across different data sources, properties testable from data that have been exploited recently, for example [133, 64, 143].

Assumption 1. We assume that we have access to input and output pairs (X,Y) observed across heterogeneous data sources or environments e, defined as a probability distribution P_e over an observation space $\mathcal{X} \times \mathcal{Y}$ that arises, just like new unseen data, from manipulations in the distribution of exogenous variables in an underlying model $\mathcal{M}$.

Assumption 2. For the remainder of this section *only*, consider restricting ourselves to data sources emerging from manipulations in exogenous E_X (i.e. manipulations in observed variables) in an underlying additive noise model.

It may be shown then, by considering the distributions of error terms $Y - f \circ \phi(X)$ and its correlation with any function of X, that the inner product $\nabla_\beta f(z;\beta_0)\varepsilon$, even if **non-zero** due to unobserved confounding as shown in (5.3) it converges to a **fixed unknown value equal across training environments**. Please see the Appendix for a precise statement of the assumptions and context.

Proposition 1 (Derivative invariance). *For any two environment distributions P_i and P_j generated under assumption 2, it holds that, up to disturbance terms, the causal parameter β_0*

satisfies,

$$\mathop{\mathbb{E}}_{(x,y)\sim P_i} \nabla_\beta f(z;\beta_0)(y - f(z;\beta_0)) - \mathop{\mathbb{E}}_{(x,y)\sim P_j} \nabla_\beta f(z;\beta_0)(y - f(z;\beta_0)) = 0. \tag{5.4}$$

Proof. All proofs are given in the Appendix.

This *invariance* across environments must hold for causal parameters (under certain conditions) *even* in the presence of unobserved confounders.

5.1.3 Remarks

A few remarks are necessary concerning this relationship and its extrapolation properties.

- The first is based on the observation that, up to a constant, each inner product in (5.4) is the gradient of the squared error with respect to β. This reveals that the optimal predictor, in the presence of unobserved confounding, is not one that produces minimum loss but one that produces a *non-zero* loss gradient *equal* across environments.

  Therefore, seeking minimum error solutions, even in the population case, produces estimators with *necessarily* unstable correlations because the variability due to unobserved confounders is not explainable from observed data. Forcing gradients to be zero then *forces* models to utilize artifacts of the specific data collection process that are not related to the input-output relationship; and, for this reason, will not in general perform outside training data.

- From (5.4) we may pose a sequence of moment conditions for each pair of available environments. We may then seek solutions β that make all of them small simultaneously. Solutions are unique if the set of moments is sufficient to identify $\beta^\star$ exactly (and given our model assumptions may be interpreted as causal and robust to certain interventions).

  In the Appendix, we revisit our introductory example to show that indeed this is the case, and that other invariances exploited for causality and robustness (such as [5, 96]) do not hold in the presence of unobserved confounding and give biased results.

- In practice of course only a *set* of solutions may be identified with the moment conditions in Proposition 1 with no performance guarantees for any individual solutions, and no guarantees if assumptions fail to hold.

  Moreover, even if accessible, we have seen in Figure 5.1 that causal solutions may not always be desirable under more general shifts (for example shifts in unobserved variables).

5.2 A Robust Optimization Perspective

In this section we motivate a relaxation of the ideas presented using the language of robust optimization.

One strategy is to optimize for the worst case loss across environments which ensures accurate prediction on any convex mixture of training environments [14]. The space of convex mixtures, however, can be restrictive. For instance, in high-dimensional systems perturbed data is likely occur at a new vertex not represented as a linear combination of training environments. We desire performance guarantees outside this convex hull.

We consider in this section problems of the form of (5.1) over an *affine* combination of training losses, similarly to [96], and show that they relate closely to the invariances presented in Proposition 1.

Let $\Delta_\eta := \{\{\alpha_e\}_{e\in\mathcal{E}} : \alpha_e \geq -\eta, \sum_{e\in\mathcal{E}} \alpha_e = 1\}$ be a collection of scalars and consider the set of distributions defined by $\mathcal{P} := \{\sum_{e\in\mathcal{E}} \alpha_e P_e : \{\alpha_e\} \in \Delta_\eta\}$, all affine combinations of distributions defined by the available environments. $\eta \in \mathbb{R}$ defines the strength of the extrapolation, $\eta = 0$ corresponds to a convex hull of distributions but above that value the space of distributions is richer, going beyond what has been observed: affine combinations amplify the strength of manipulations that generated the observed training environments. The following theorem presents an interesting upperbound to the robust problem (5.1) with affine combinations of errors.

Theorem 1 *Let $\{P_e\}_{e\in\mathcal{E}}$, be a set of available training environments. Further, let the parameter space of β be open and bounded. Then, the following inequality holds,*

$$\sup_{\{\alpha_e\}\in\Delta_\eta} \sum_{e\in\mathcal{E}} \alpha_e \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}(f\circ\phi(x),y) \leq \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}(f\circ\phi(x),y)$$
$$+ (1+n\eta)\cdot C\cdot \left|\left| \sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \nabla_\beta \mathcal{L}(f\circ\phi(x),y) - \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \nabla_\beta \mathcal{L}(f\circ\phi(x),y) \right|\right|_{L_2},$$

where $||\cdot||_{L_2}$ denotes the L_2-norm, C is a constant that depends on the domain of β, $n := |\mathcal{E}|$ is the number of available environments and $e\sim\mathcal{E}$ loosely denotes sampling indeces with equal probability from $\mathcal{E}$.

Interpretation. This bound illustrates the trade-off between the invariance of Proposition 1 (second term of the inequality above) and prediction in-sample (the first term). A combination of them upper-bounds a robust optimization problem over affine combinations of training environments, and depending how much we weight each objective (prediction versus invariance) we

can expect solutions to be more or less robust. Specifically, for $\eta = -1/n$ the objective reduces to ERM, but otherwise the upperbound increasingly weights differences in loss derivatives (violations of the invariances of section 5.1.2), and in the limit ($\eta \to \infty$) can be interpreted to be robust at least to *any* affine combination of training losses.

Remark on assumptions. Note that the requirement that $\mathbb{F}$ be fixed or Assumption 2, is not necessary for generalization guarantees. As long as new data distributions can be represented as affine combinations of training distributions, we can expected performance to be as least as good as that observed for the robust problem in Theorem 1.

5.2.1 Proposed objective

Our proposed learning objective is to guide the optimization of ϕ and β towards solutions that minimize the upperbound in Theorem 1. Using Lagrange multipliers we define the general objective,

$$\underset{\beta,\phi}{\text{minimize}} \underset{(x,y)\sim P_e, e\sim\mathcal{E}}{\mathbb{E}} \mathcal{L}\left(f\circ\phi(x), y\right) + \lambda \cdot \underset{e\sim\mathcal{E}}{\text{Var}}\left(|| \underset{(x,y)\sim P_e}{\mathbb{E}} \nabla_\beta \mathcal{L}\left(f\circ\phi(x), y\right) ||_{L_2} \right),$$

where $\lambda \geq 0$. We call this problem Derivative Invariant Risk Minimization (DIRM).

This objective shares similarities with the objective proposed in [96]. The authors considered enforcing equality in environment-specific losses, rather than derivatives, as regularization, which can also be related to a robust optimization problem over an affine combination of errors. We have seen in section 5.1.2 however that equality in losses is not expected to hold in the presence of unobserved confounders.

Remark on optimization. The L_2 norm in the regularizer is an integral over the domain of values of β and is in general intractable. We approximate this objective in practice with norms on functional evaluations at each step of the optimization rather than explicitly computing the integral. We give more details and show this approximation to be justified empirically in the Appendix.

5.2.2 Robustness in terms of interventions

As is apparent in Theorem 1, performance guarantees on data from a new environment depend on the relationship of new distributions with those observed during training.

Let $f \circ \phi_{\lambda \to \infty}$ minimize $\mathcal{L}$ among all functions that satisfy all pairs of moment conditions defined in (5.4); that is, a solution to our proposed objective in (5.5) with $\lambda \to \infty$. At optimality, it holds that gradients evaluated at this solution are equal across environments. As a consequence of Theorem 1, the loss evaluated at this solution with respect to *any* affine combination of environments is bounded by the average loss computed in-sample (denoted L, say),

$$\sum_{e \in \mathcal{E}} \alpha_e \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}(f \circ \phi(x), y) \leq L, \qquad \text{for any set of } \alpha_e \in \Delta_\eta. \tag{5.5}$$

From the perspective of interventions in the underlying causal mechanism, this can be seen as a form of data-driven predictive stability across a range of distributions whose perturbations occur in the same direction as those observed during training.

Example. Consider distributions P of a univariate random variable X given by affine combinations of training distributions P_0 with mean 0 and P_1 which, due to intervention, has mean 1 so that, using our notation, $\mathbb{E}_P X = \alpha_0 \mathbb{E}_{P_0} X + \alpha_1 \mathbb{E}_{P_1} X$, $\alpha_0 = 1 - \alpha_1 \geq -\eta$. $\mathbb{E}_P X \in [-\eta, \eta]$ and thus we may expect DIRM to be robust to distributions subject to interventions of magnitude $\pm\eta$ on X and any magnitude in the limit $\eta \to \infty$ (or equivalently $\lambda \to \infty$). With this reasoning, however, note that the "diversity" of training environments has a large influence on whether we can interpret solutions to be causal (for which we need interventions on all observed variables and unique minimizers) and robustness guarantees: for instance, with equal means in P_0 and P_1 affine combinations would not extrapolate to interventions in the mean of X. This is why we say that interventions in test data must have the same "direction" as interventions in training data (but interventions can occur on observed, unobserved or target variables).

Using our simple example in Figure 5.1 to verify this fact empirically, we consider 3 scenarios corresponding to interventions on exogenous variables of X, H and Y. In each, training data from two environments is generated with means in the distribution of the concerned variables set to a value of 0 and 1 respectively, everything else being equal ($\sigma^2 := 1, H := E_H \sim \mathcal{N}(0,1)$). Performance is evaluated on out-of-sample data generated by increasing the shift in the variable being studied up to a mean of 5. In all cases, we see in Figure 5.2 that performance is stable to increasing perturbations in the system as long as the heterogeneity in the data allows us to capture the direction of the unseen shift.

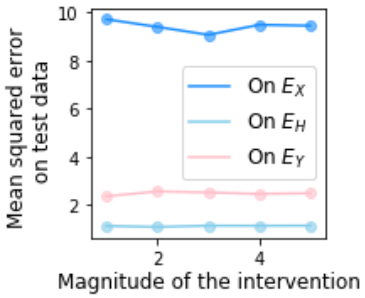


Figure 5.2: Stability to general shifts.

5.2.3 Stability of certain optimal solutions

A special case may also be considered when the underlying system of variables and the available environments allow for optimal solutions $f \circ \phi_{\lambda \to \infty}$ and $f \circ \phi_{\lambda=0}$ to coincide. In this case, the learned representation $\phi(x)$ results in a predictor f optimal on average *and* simultaneously with equal gradient in each environment, thus,

$$|| \mathop{\mathbb{E}}_{(x,y)\sim P_e} \nabla_\beta \mathcal{L}(f \circ \phi(x), y) ||_{L_2} = 0, \qquad \text{for all } e \in \mathcal{E}. \tag{5.6}$$

For this representation ϕ, it follows that optimal solutions f learned on any new dataset sampled from an affine combination of training distributions coincides with this special solution. This gives us a sense of reproducibility of learning. In other words, if a specific feature is significant for predictions on the whole range of λ with the available data then it will likely be significant on new (related) data.

Contrat with IRM. The above special case where all solutions in our hyperparameter range agree has important parallels with IRM [5]. The authors proposed a learning objective enforcing representations of data with minimum error on average and across environments, such that at optimum $\mathbb{E}_{P_i} Y|\phi^\star(X) = \mathbb{E}_{P_j} Y|\phi^\star(X)$ for any pair $(i, j) \in \mathcal{E}$. With unobserved confounding, both learning paradigms agree but, with unobserved confounding, minimum error solutions of IRM by design converge to spurious associations (see the discussion after equation (5.4)) and are not guaranteed to generalize to more general environments. For example, in the presence of additive unobserved confounding H, irrespective of ϕ, we may have $\mathbb{E}_{P_i} Y|\phi^\star(X) = \phi^\star(X) + \mathbb{E}_{P_i} H \neq \phi^\star(X) + \mathbb{E}_{P_j} H = \mathbb{E}_{P_j} Y|\phi^\star(X)$ if the means of H differ. The sought invariance then does not hold.

5.3 Related work

There has been a growing interest in interpreting shifts in distribution to fundamentally arise from interventions in the causal mechanisms of data. Peters et al. [133] exploited this link for causal inference: causal relationships by definition being invariant to the observational regime. Invariant solutions, as a result of this connection, may be interpreted also as robust to certain interventions [117], and recent work has explored learning invariances in various problem settings [5, 143, 96, 65]. Among those, we note the invariance proposed in [143], the authors seek to recover causal solutions with unobserved confounding. Generalization properties of these solutions were rarely studied, with one exception being Anchor regression

[144]. The authors proposed to interpolate between empirical risk minimization and causal solutions with explicit robustness to certain interventions in a linear model. In a related thread, Bulhman et al. [118] considered potential shifts in the actual mechanisms relating cause and effect and, maximize instead for relative performance over the worst varying-coefficient model. The present work may be interpreted as a non-linear formulation of this principle with a more general study of generalization.

Domain generalization represent one direction of out-of-sample generalization by explicitly learning representations projecting out superficial environment-specific information. Recent work on domain generalization has included the use data augmentation [184, 160], meta-learning to simulate domain shift [106] and adversarially learning representations that are environment invariant [61], even though explicitly aligning representations has important caveats when label distributions differ, articulated for instance in [5]. Distributionally robust optimization is a first related line of research that explicitly solves a worst-case optimization problem (5.1). A popular approach is to define $\mathcal{P}$ as a ball around the empirical distribution $\hat{P}$, for example using f-divergences or Wasserstein balls of a defined radius [97, 53, 55, 162, 189, 1, 54]. These are general and multiple environments are not required, but this also means that sets are defined agnostic to the geometry of plausible shifted distributions, and may therefore lead to solutions, when tractable, that are overly conservative or do not satisfy generalization requirements [55]. Transfer learning is a second related line of research that considers data from both training and testing domains (and possibly labeled in the testing domain) to be available, the challenge being how to most efficiently use training data to learn optimal hypotheses in the testing domain. Bounds on the error in estimation have been developed in that setting using a measure of discrepancy between training and testing distributions [44, 13] and consistent estimation may be achieved if additional structure (e.g. covariate shift, label shift, etc.) can be assumed [29, 116].

5.4 Experiments

Data linkages, electronic health records, and bio-repositories, are increasingly being collected to inform medical practice. As a result, also prediction models derived from healthcare data are being put forward as potentially revolutionizing decision-making in hospitals. Recent studies [28, 180], however, suggest that their performance may reflect not only their ability to identify disease-specific features, but also their ability to exploit spurious correlations due to unobserved confounding (such as varying data collection practices): a major challenge for the reliability of

decision support systems. In this section, we explore this pattern conducting a wide analysis of domain generalization on image, speech and tabular data from the medical domain.

We consider the following baseline algorithms for performance comparisons.

- Empirical Risk Minimization (**ERM**) that optimizes for minimum loss agnostic of data source.
- Domain Robust Optimization (**DRO**) [149] that optimizes for minimum loss across the worst convex mixture of training environments.
- Domain Adversarial Neural Networks (**DANN**) [61] that use domain adversarial training to facilitate transfer by augmenting the neural network architecture with an additional domain classifier to enforce the distribution of $\phi(X)$ to be the same across training environments.
- Invariant Risk Minimization (**IRM**) [5] that regularizes ERM ensuring representations $\phi(X)$ be optimal in every observed environment.
- Risk Extrapolation (**REx**) [96] that regularizes for equality in environment losses instead of considering their derivatives.

All trained models use the same convolutional or fully-connected architecture, where appropriate. Performance results are given in Table 5.1. Further experimental details and pseudo-code for DIRM can be found in Appendix D.4.

5.4.1 Diagnosis of Pneumonia with chest X-ray data

In this section, we attempt to replicate the study in [195]. The authors observed a tendency of image models towards exploiting spurious correlations for the diagnosis on pneumonia from patient Chest X-rays that do not reproduce outside of training data. We use publicly available data from the National Institutes of Health (NIH) [188] and the Guangzhou Women and Children's Medical Center (GMC) [92]. Differences in distribution are manifest, and can be seen for example in the top edge of mean pneumonia-diagnosed X-rays shown in Figure 5.3.

Table 5.1: Accuracy of predictions in percentages (%). Uncertainty intervals are standard deviations. All datasets are approximately balanced, 50% performance is as good as random guessing.

	Pneumonia Prediction		**Parkinson Prediction**		**Survival Prediction**	
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**
ERM	91.4 (± .7)	52.4 (± 1)	95.7 (± .5)	62.9 (± 1)	93.2 (± .4)	75.3 (± .9)
DRO	91.2 (± .5)	53.1 (± .6)	94.0 (± .3)	69.8 (± 2)	90.5 (± .4)	75.5 (± .8)
DANN	92.3 (± 1)	57.1 (± 2)	91.6 (± 2)	51.4 (± 5)	89.3 (± .8)	73.9 (± .9)
IRM	89.5 (± 1)	58.6 (± 2)	93.6 (± 1)	71.4 (± 2)	91.9 (± .6)	75.7 (± .8)
REx	87.7 (± 1)	57.9 (± 2)	92.0 (± 1)	72.4 (± 2)	91.3 (± .5)	75.0 (± .9)
DIRM	84.3 (± 1)	63.7 (± 3)	93.1 (± 2)	72.8 (± 2)	91.4 (± .6)	77.9 (± 1)

In this experiment, we exploit this (spurious) pathology correlation to demonstrate the need for solutions robust to changes in site-specific features. We construct two training sets, in each case 90% and 80% of pneumonia-diagnosed patients were drawn from the NIH dataset and the remaining 10% and 20% of the pneumonia-diagnosed patients were drawn from the GMC dataset; the reverse logic (10%/90% split) was followed for the test set.

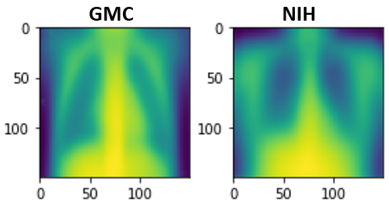


Figure 5.3: Average pneumonia X-ray.

Our results show that DIRM outperforms, suggesting that the proposed invariance guides solutions better towards robustness even to changes due to unobserved factors.

5.4.2 Diagnosis of Parkinson's disease with voice recordings

Parkinson's disease is a progressive nervous system disorder that affects movement. Symptoms start gradually, sometimes starting with a barely noticeable tremor in a patient's voice. This section investigates the performance of predictive models for the detection of Parskinson's disease, trained on voice recordings of vowels, numbers and individual words and tested on vowel recordings of unseen patients.

We used the UCI Parkinson Speech Dataset with given training and testing splits [150]. Even though the distributions of features will differ in different types of recordings and patients, we would expect the underlying patterns in speech to reproduce across different samples. However, this is not the case for correlations learned with baseline training paradigms (Table 5.1). This suggests that spurious correlations due to the specific type of recording (e.g. different vowels or numbers), or even chance associations emphasized due to low sample sizes (120 examples), may be responsible for poor generalization performance. Our results show that correcting for

spurious differences between recording types (DIRM, IRM, REx) can improve performance substantially.

5.4.3 Survival prediction with electronic health records

This section investigates whether predictive models transfer across data from different medical studies – the MAGGIC studies [115] – all containing patients that experienced heart failure. The problem is to predict survival within 3 years of experiencing heart failure from a total of 33 demographic variables. We introduce a twist however, explicitly introducing unobserved confounding by omitting certain predictive variables. The objective is to test performance on new studies with *shifted* distributions, while knowing that these occur predominantly due to variability in unobserved variables.

Confounded data is constructed by omitting a patient's age from the data, found in a preliminary correlation analysis to be associated with the outcome as well as other significant predictors such as blood pressure and body mass index. This example is constructed to be able to control for how unobserved variables shift but note that we can expect similar phenomena in many other scenarios, where for instance a prediction model is taken to patients in a different hospital or country with fundamental shifts in the distribution of very relevant variables (e.g. socio-economic status, ethnicity, diet, etc.) even though this information is not reported in the data. Performance is tested on all studies of over 500 patients with balanced death rates, each having slightly different age distributions. (We give more details in Appendix D.4). We found DIRM, robust to changes in unobserved variables, to outperform all other methods.

Influential variables that reproduce across datasets. In the following, we tackle the problem of *reproducibility* of learned influential features across different experiments. Reproducing conclusions of influential features in different studies with potential shifts in the distribution is an important challenge, especially in healthcare where heterogeneity between patient populations is high. We showed in section 5.2.3 that in the event that the optimal predictor is invariant as a function $\lambda \in [0, \infty)$, optimal predictors estimated in *every* new dataset in the span of observed distributions, should be *stable*.

We consider here a form of diluted stability for feature selection. For a single layer network, we consider significant those covariates with estimated parameters bounded away from zero in all solutions in the range $\lambda \in [0, 1]$. Comparisons are made with ERM (conventional logistic regression), both methods trained separately on 100 random pairs of studies. Figure 5.4 shows how many features (in the top 10 of predictive features) from each model intersect across pairs of studies. In constrast to ERM, our objective recovers significant features much more consistently.

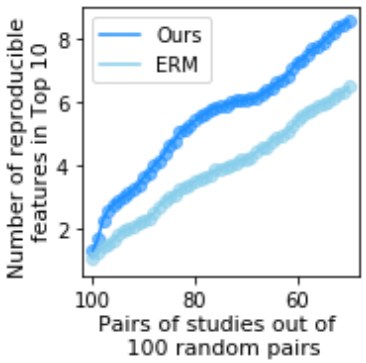


Figure 5.4: Reproducible features.

5.5 Conclusions

We have studied the problem of out-of-sample generalization from a new perspective, grounded in the underlying causal mechanism generating new data that may arise from shifts in observed, unobserved or target variables. Our proposal is a new objective that is provably robust to certain shifts in distribution, and is informed by certain statistical invariances in the presence of unobserved confounders. Our experiments show that we may expect better generalization performance and also better reproducibility of influential features in problems of variable selection.

A limitation of our approach is that robustness guarantees crucially depend on the (unobserved) properties of available data. Using the proposed approach, Derivative Invariant Risk Minimization for prediction generally does not guarantee protection against unsuspected events. More specifically, we can not expect robust prediction when the heterogeneity in test data sets is different from the restricted set of shift interventions that have been observed on the training data sets. For example, in Theorem 1, the supremum contains distributions that lie in the affine combination of training environments, as opposed to arbitrary distributions.

Chapter 6

Scoring DAGs with Dense Unobserved Confounders

Unobserved confounding is one of the greatest challenges for causal discovery. The case in which unobserved variables have a potentially widespread effect on many of the observed ones is particularly difficult because most pairs of variables are conditionally dependent given any other subset. In this chapter we show that beyond conditional independencies, unobserved confounding in this setting leaves a characteristic footprint in the observed data distribution that allows for disentangling spurious and causal effects. Using this insight, we demonstrate that a sparse linear Gaussian directed acyclic graph among observed variables may be recovered approximately and propose an adjusted score-based causal discovery algorithm that may be implemented with general purpose solvers and scales to high-dimensional problems. We find, in addition, that despite the conditions we pose to guarantee causal recovery, performance in practice is robust to large deviations in model assumptions.

Introduction

Unmeasured confounding is a long-standing challenge for reliably drawing causal inferences from observational data. This is because, in the presence of unobserved confounding, correlations observed in data are compatible with many potentially contradictory causal explanations, leaving the scientist unable to distinguish between them.

The problem of causal discovery is to define constraints in the data distribution (e.g. conditional independencies) to infer the causal graph [129]. In the presence of unobserved confounding, one popular way forward has been to seek an *equivalence class* of mixed graphical models,

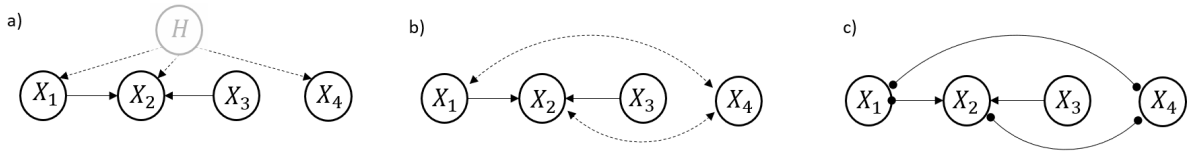


Figure 6.1: a) A DAG with unobserved confounding; b) the corresponding MAG; c) the equivalence class of MAGs representing the same conditional independences as the DAG (dots • indicate undetermined causal direction). This example considers observed variables X_1, X_2, X_3 and X_4, confounded by H which is unobserved. In genetics, these may represent the activity of a set of genes, confounded by the amount of ozone in the air or patient variation over time [an example from 60]. Due to H most pairs of variables are conditionally dependent given any subset of other variables and thus few edges in the inferred equivalence class can be oriented.

called maximal ancestral graphs (MAGs), including directed, bidirected and undirected edges representing different types of possible causal dependencies compatible with observed conditional independencies [139]. This approach is compelling because it requires no assumptions on the functional relationships between variables or even knowledge on the number or type of unobserved confounders to consistently identify equivalence classes [163, 42, 40, 176, 178].

In some problems however, equivalence classes are largely uninformative as to the underlying causal relationships between observed variables. In genetics for example, gene expression measurements are often confounded by batch effects, degradation and other specifics of the experiment, leaving most pairs of gene expression measurements conditionally dependent given any subset of other measurements [60, 102]. A similar pattern occurs in finance with asset prices driven by a common political climate or exogenous shocks even though these events are often not explicitly recorded in data [34]. In these examples, graphically, unobserved confounding, when *dense* in its effect on observables (i.e. unobserved variables having an effect on *many* of the observed ones), leaves most edges in the equivalence class of MAGs undetermined. See Figure 6.1 for a concrete example.

In this context, we show that we can make progress by restricting ourselves to learning the directed edges among observed variables in a causal MAG (i.e. a directed acyclic graph (DAG)). We study the setting of a high-dimensional system of variables $X \in \mathbb{R}^p$, in an underlying linear model whose causal interactions are specified by the non-zero entries of a sparse adjacency matrix $W \in \mathbb{R}^{p \times p}$, in the presence of *dense* unobserved confounding $H \in \mathbb{R}^q$,

$$X = WX + BH + E, \tag{6.1}$$

where E is an independent vector of errors but realizations of X may be confounded by H through $B \in \mathbb{R}^{p \times q}$.

Contributions. A practical consequence of dense unobserved confounding is that the contribution of the matrix of confounded contributions B to the matrix of observed correlations $\mathrm{Cov}(X)$ is *different* (in a characteristic sense) from the contribution due to the matrix of causal contributions W. We can then adjust for confounded contributions by analogizing DAG learning to a regression problem involving a sparse plus dense superposition of matrices [31, 158, 33], in this case interpreted as causal and confounded contributions respectively in the context of unobserved confounding.

Specifically, we show that one can formulate DAG learning among p observed variables in the presence of dense unobserved confounding as the solution of an optimization program:

$$\text{minimize } \mathcal{S}(W; \mathbf{X}) \quad \text{such that} \quad W \in \mathbb{D}, \tag{6.2}$$

where $\mathbb{D}$ is the set of $p \times p$ matrices representing the weighted adjacency matrix of a DAG and $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the data. $\mathcal{S}$ is known as the score function. Estimators of this form have a long history in causal discovery [4, 153, 37, 201, 27, 110], predominantly in the fully observed setting.

Our contributions are three-fold.

1. We show that in high-dimensional ($p \gg n$) linear models (6.1) the spectrum of the confounded data matrix is characteristically different than would be expected without unobserved confounding.

2. With this insight, we propose a score function $\mathcal{S}$ and problem (6.2) whose solution has explicit finite-sample true positive guarantees.

3. We develop a practical two-stage algorithm, the Deconfounded Score (DECS) method, leveraging standard gradient-based optimization solvers and algebraic acyclicity formulations of DAGs that has the practical benefits of being much simpler and scaling better to large samples and high-dimensional feature spaces than alternative independence-based approaches.

6.1 Related work

This chapter primarily engages with the literature on causal discovery in the presence of unobserved confounding but also draws on insights from high-dimensional linear regression and factor models.

We argue for exploiting properties of the spectrum of data matrix to recover a causal DAG among observed variables in high-dimensional systems. We contrast this approach with work that seeks conditional independencies as a route to causality, first presented in [163] and subsequently widely extended and applied e.g., [42, 40, 41]. The authors developed theoretically consistent algorithms for recovering an equivalence class of MAGs [139] which may be linked to the underlying causal structure with an assumption of faithfulness. Examples include the FCI, FCI+, RFCI and other variants that use (a polynomial number of) conditional independence tests to iteratively recover the skeleton and some edge orientations. A second class of algorithms instead propose to search greedily for graphs optimizing a score function defining goodness of fit on the observed data. For instance, [176] proposed a greedy search algorithm maximizing a penalized Gaussian likelihood score over the class of MAGs, [17] proposed a greedy search over partial orderings of the variables, [58] use a decomposition of the covariance matrix extending the GES algorithm [37] to unobserved confounding, [178] proposed a hybrid combination of score and independence-based algorithms, among others that consider bow-free acyclic graphs (a special case of MAGs) [124, 52].

We share the objective of seeking a consistent score function but instead aim to recover a DAG among observed variables only and do so focusing on high-dimensional spaces from a penalized regression perspective, relying instead on the principle of independent conditionals [82] to link the spectrum of the data matrix to causality. This challenge is related to the literature on identifiability in high-dimensional regression [36, 31, 158, 33] and estimation in linear factor models [56, 57, 6, 57, 23]. For instance, in different variations of the underlying factor model it is possible to consistently recover a decomposition of regression parameters or covariance matrices into a sparse component and a dense or low-rank component separately. This chapter applies this theory to extend (fully-observed data) score-based DAG learning consistency results (e.g. [4, 3]) to a special case of unobserved confounding that could not be consistently analysed before.

6.2 Problem formulation

We use the language of structural causal models (SCMs) as our basic semantical framework to enable us to specify functional relationships between variables.

We suppose a structural causal model describes a natural phenomenon of interest, partially observed through a random vector $X = (X_1, \ldots, X_p)$ satisfying,

$$X = WX + BH + E, \tag{6.3}$$

where $W \in \mathbb{D}$ specifies the causal variable relationships in X. $H = (H_1, \ldots, H_q)$ is a vector of q unobserved Gaussian confounders that influence X through a dense[1], fixed matrix $B \in \mathbb{R}^{p \times q}$. $E = (E_1, \ldots, E_p)$ a vector of independent sources of noise, also drawn from a Gaussian distribution. We will denote $\mathbf{X} \in \mathbb{R}^{n \times p}$ as the data matrix, each row independently sampled from (6.3) and we will assume $p \gg n$. W defines a DAG over the observed variables, if $[W]_{ij} \neq 0$ we will say that $X_j \in \mathtt{Pa}(X_i)$ is a causal parent of X_i.

Our goal is to define a score function $\mathcal{S} : \mathbb{R}^{p \times p} \to \mathbb{R}$, involving only the observed data, that provably attains a minimum at the weighted adjacency matrix W of the underlying DAG.

6.2.1 The challenge of high-dimensional data

In high-dimensional systems, defining a function that scores candidate adjacency matrices W is intrinsically ill-posed without further structure. When $\mathrm{Rank}(\mathbf{X}) < p$, e.g. when $p > n$, there are infinitely many solutions with minimum score. Given one solution $W^\star$, the quantity $W^\star + \varepsilon$ is also a solution for any ε in the null space of $\mathbf{X}$. Moreover, even if only signs are desired in the underlying DAG (i.e. we seek to know whether each estimated causal effect raises or lowers the probability of outcomes in children nodes), this type of non-uniqueness makes interpretation of solutions cumbersome: for any i and at least one $j \in \{1, \ldots, p\}$, we will have $W^\star_{ij} > 0$ for one solution, but $W^\star_{ij} < 0$ for another solution. Constraining solutions to be sparse (i.e. few edges in comparison to the number of variables involved) is one way to overcome this problem [76].

[1] e.g. we may consider B dense if all entries are drawn from a Gaussian distribution

6.2.2 The challenge of confounded data

An assumption of sparsity on solutions to score-based optimization problems such as (6.2) is not appropriate however. The inferred matrix of associations will typically be dense as a result of confounding. We may write for instance,

$$X = (W + C)X + (BH - CX) + E, \tag{6.4}$$

where $C \in \mathbb{R}^{p \times p}$ is chosen such that $\mathrm{Cov}(BH - CX, X) = 0$ (interpreted as the covariance of every pair of random variables in each of the two vector arguments to be zero). C is the scaled projection of H on X: $C = \mathrm{Cov}(X)^{-1}\mathrm{Cov}(X,H)B$, and represents the bias introduced in the estimation of W due to the contributions of unobserved confounding variables H. If we ignore confounding, we shall have $W + C$ as the target of score-based algorithms instead of W. And as mentioned in section 6.2.1, it will typically not be accessible in high-dimensional systems. The bias in estimation of W is potentially large if $||\mathbf{X}C||$ is large. We rename the error vector of this model $\tilde{E} := (BH - CX) + E$, each entry independently distributed and independent of X by construction.

6.3 Adjusted scoring of DAGs

In this section, we describe the principle of independent causal mechanisms which motivates an adjusted score function that mitigates the contribution of unobserved confounding while preserving the causality among observed variables.

6.3.1 The asymmetry of confounding

Confounded relationships between observed variables leave a characteristic statistical footprint in the data distribution.

If we were to be given the underlying causal structure and all variables fully observed ($H = 0$), we could write,

$$\mathbb{P}(X_1, \ldots, X_p) = \prod_{i=1}^{p} \mathbb{P}(X_i | \mathtt{Pa}(X_i)), \tag{6.5}$$

where each $\mathbb{P}(X_i|\texttt{Pa}(X_i))$ denotes the conditional distribution of X_i on its parents $\texttt{Pa}(X_i)$. The conditional distributions have the property of describing an invariant mechanism of nature (e.g. in the same manner that physical laws are invariant to location and time), that should be independent of the distribution of the causes $\mathbb{P}(\texttt{Pa}(X_i))$ [82, 83]. Importantly, this independence does not hold in the presence of unobserved confounders, since these variables induce a correlation between X_i and its parents $\texttt{Pa}(X_i)$.

Given that the underlying model of variable associations (6.3) is linear we may define this independence criterion by associating each $\mathbb{P}(X_i|\texttt{Pa}(X_i))$ with the set of parameters $\mathbf{w}_i$ (i.e. the i^{th} row of W). Following this reasoning, $\mathbf{w}_i$ should be independent from the distribution of its parents $\mathbb{P}(\texttt{Pa}(X_i))$, in our case fully specified by the matrix of second moments of X. Specifically, it would be unexpected to find $\mathbf{w}_i$ aligned in any specific manner to large principal components of $\mathbf{X}$ (i.e. large eigenvalues of $\text{Cov}(\mathbf{X})$).

In the presence of unobserved confounding, this changes, independence of causal mechanisms is not expected to hold and will induce a statistical footprint in the distribution of the observed data that is different than it should be without confounding. We will see that this holds specifically in our model in the next subsection.

6.3.2 Adjusting for confounding

We have mentioned that the bias in estimation of W is potentially large if $||\mathbf{X}C||$ is large in (6.4), but the *direction* of this contribution tends to be concentrated in specific vectors related to the covariance matrix of X. Specifically, assuming $\text{Cov}(H) = \text{Cov}(E) = I$ and W sparse, each column in B in (6.3) tends to be *approximately* aligned with $\text{Cov}(X) = (I-W)^{-1}(BB^T + I)(I-W)^{-T}$ since each column in B is an eigenvector of $(BB^T + I)$ with large eigenvalue (all other eigenvectors have a corresponding eigenvalue equal to 0).

And therefore also the direction of the transformation $C = \text{Cov}(X)^{-1}\text{Cov}(X,H)B$, as a multiple of B, must be aligned with large eigenvectors of $\text{Cov}(X)$. Under the principle of independent conditionals, such dependence or alignment between coefficients W of X (that define the causal contribution) and its distribution is unlikely. We can expect therefore that shrinking large principal components of $\mathbf{X}$ shrinks the contribution of C in our estimates but leaves the contribution due to causal coefficients unchanged – as these are largely orthogonal to the direction of large principal components of $\mathbf{X}$.

One approach, originating in [56] in the context of high-dimensional regression, is thus to remove or truncate large singular values of $\mathbf{X}$ leaving the direction of singular vectors unchanged.

Let $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$ be the singular value decomposition of $\mathbf{X}$, where $\mathbf{U} \in \mathbb{R}^{n\times r}$, $\mathbf{D} \in \mathbb{R}^{r\times r}$, $\mathbf{V} \in \mathbb{R}^{p\times r}$, and where $r = \min(n,p)$ is the rank of $\mathbf{X}$. We write $d_1 \leq d_2 \leq \cdots \leq d_r$ for the diagonal elements of $\mathbf{D}$. We use the truncated form of the singular value decomposition, which uses only non-zero singular values.

We define the adjusted matrix $\tilde{\mathbf{X}} := \mathbf{F}\mathbf{X}$ as a transformation of $\mathbf{X}$ by $\mathbf{F} \in \mathbb{R}^{n\times n}$ that upper-bounds each singular value to $\tilde{d} := \text{median}(d_1,\ldots,d_r)$: $\mathbf{F} := \mathbf{U}\tilde{\mathbf{D}}\mathbf{U}^\intercal$, where $\tilde{\mathbf{D}}$ is diagonal with each element on the diagonal equal to $[\tilde{\mathbf{D}}]_{ii} := \min(d_i, \tilde{d})/d_i$. As demonstrated in [33] for this transformation matrix, without adjustments, the contribution due to unobserved confounding $||\mathbf{X}C||$ may be as large as $||\mathbf{H}B||$ which can be shown to be of the order of $p\sqrt{n}||B||$ while $||\tilde{\mathbf{X}}C||$ is of the order of $p||B||$, which is much smaller.

6.3.3 An adjusted score function

A score-based DAG estimator that derives from this approach is immediate, formulated as the solution of a constrained optimization problem,

$$\widehat{W} \in \underset{W\in\mathbb{D}}{\text{argmin}} \quad \mathcal{S}(W;\mathbf{X}),$$
$$\mathcal{S}(W;\mathbf{X}) := \frac{1}{2n}||\tilde{\mathbf{X}} - \tilde{\mathbf{X}}W||_F^2 + \lambda||W||_1, \tag{6.6}$$

where $\tilde{\mathbf{X}} = F\mathbf{X} \in \mathbb{R}^{n\times p}$ is the linear transformation of the data matrix that truncates large principal components. $\mathcal{S}$ is the penalized mean squared score function in Frobenius norm that quantifies how well W agrees with observations, $\lambda||W||_1$ is the scaled sum of the magnitude of the entries in W, and $\lambda > 0$.

6.3.4 A guarantee on recovery of W

We are motivated by the prior that large principal components should align more closely with correlations driven by unobserved confounding. An important question is whether solutions to the adjusted optimization problem in fact converge, and if so, whether they converge to the underlying causal structure W.

The problem in (6.6) can be interpreted as optimization over a family of neighbourhood regression problems, each variable regressed on its non-descendants. This decomposition can be used to derive uniform bounds on recovery error. In particular, [4] first showed that imposing

sparsity on the true DAG substantially reduces the number of regressions, otherwise equal to $2^{p-1}p$ (since the topological ordering of the DAG, or the set of non-descendants for each variable is unknown a priori) and intractable in general. Penalized score-based learning without unobserved confounding, they showed, efficiently and provably recovers a sparse DAG $W_{\min}$ with minimum conditional variance, also called minimum-trace DAG. If unique $W_{\min}$ equals W, otherwise there is technically no truth to approximate from data, though penalized score-based learning does converge to a sparse representative among the class of minimum-trace DAGs. We refer to [4, 3] for more details.

In this section, we show that a similar strategy applies in our setting, with the difference however that each neighbourhood regression problem, instead of being a penalized regression problem, is formulated as the following adjusted, penalized regression problem,

$$\underset{\mathbf{w}_i \in \mathbb{R}^p,\ \text{supp}(\mathbf{w}_i) \subset S}{\arg\min} \quad \frac{1}{2n} ||\tilde{\mathbf{X}}_i - \tilde{\mathbf{X}}\mathbf{w}_i||_2^2 + \lambda ||\mathbf{w}_i||_1. \tag{6.7}$$

S is a subset of all variables other that X_i that defines a *neighbourhood* of X_i. $\tilde{\mathbf{X}}_i \in \mathbb{R}^n$ is the i^{th} column of $\tilde{\mathbf{X}}$, $\mathbf{w}_i \in \mathbb{R}^p$ is the i^{th} column of W (i.e. the regression parameters defining the parents of X_i) and $\text{supp}(\mathbf{w}_i)$ denotes the support of $\mathbf{w}_i$.

To obtain uniform bounds on the error in DAG estimation as in [4, 3] it suffices to show that each regression parameter $\mathbf{w}_i$ can be recovered consistently. Bounds on the estimation of $\mathbf{w}_i$ (in l_1 of l_2 norms for example), exist in the high-dimensional regression literature once we recognise $\mathbf{w}_i$ as the sparse component in a sparse plus dense superposition of regression parameters [e.g. Theorem 1 in 33]. Two conditions are needed for these bounds. First, we must ensure that the perturbation C due to unobserved confounding on each individual estimated parameter is not too large (specifically, a restriction on the order of the largest singular value of $\text{Cov}(X,H)$ which holds as a consequence of confounding being dense). Second, we must ensure the transformation F to be well-behaved, i.e. not shrink the causal signal too much (specifically imposing a smallest restricted eigenvalue condition on the covariance matrix of $\tilde{X}$) but consistently lower large singular vectors of $\tilde{X}$. We refer to Appendix E.1 for a formal statement of these regularity conditions.

For any $A \in \mathbb{R}^{p\times p}$, let $\tau(A) := \min\{|a_{ij}| : a_{ij} \neq 0\}$. The quantity $\tau(W_{\min})$ measures the smallest nonzero weight in $W_{\min}$, which is a measure of the signal strength in the problem. Denote $a \gtrsim b$ to mean that $a \geq C \cdot b$ for some constant $C > 0$, and $\sigma := \max_i(\sigma_i)$ where σ_i in the standard deviation of adjusted error terms $\tilde{E}_i F$ ($\tilde{E}_i$ defined in section 6.2.2). The following Theorem

shows that the support of the minimum-trace DAG, i.e. the true edges in the underlying DAG, is contained in the support of the estimated DAG with high-probability.

Theorem. *Under regularity conditions and $W_{\min}$ unique, for $n \gtrsim s \log p$, $\lambda \gtrsim \sigma\sqrt{\log p/n}$, and $\tau(W_{\min}) \gtrsim \lambda$,*

$$\text{supp}(W_{\min}) \subseteq \text{supp}(\widehat{W}), \tag{6.8}$$

with probability $1 - \mathcal{O}(e^{k \log p})$, where k is the maximum in-degree of $W_{\min}$, i.e. the maximum number of directed edges that point into any observed node, and s is the size of the support of $W_{\min}$.

Proof. The proof is given in Appendix E.1.

Despite the presence of unobserved confounding, this result guarantees not to miss any causal edges in the true network but we may (typically) have too many false positive selections in the estimated DAG. This result is equivalent to the property of *variable screening* of the lasso estimator. Results exist also to guarantee *full* support recovery of the lasso estimator [187]. In the DAG estimation setting however, this necessitates however much stronger conditions, roughly speaking requiring that no parent of a given variable be highly correlated with "non-parent" variables, known as the incoherence condition discussed by [3].

6.3.5 Practical algorithms

This section describes a practical algorithm to solve (up to stationarity) the constrained optimization problem (6.6).

The practical challenge is to enforce efficiently the acyclicity constraint on W. One approach is to transform the traditional combinatorial optimization problem into a continuous program, using an equivalent formulation of acyclicity via the trace exponential function, due to [201]. W corresponds to an acyclic graph if and only if the function $h(W) = 0$, where $h(W) := Tr(\exp\{W \odot W\}) - p$, $\exp\{M\}$ denotes the matrix exponential of a matrix M, $\odot$ denotes the element-wise matrix product, and Tr denotes the matrix trace operator. The optimization problem becomes,

$$\underset{W \in \mathbb{R}^{d \times d}}{\text{minimize}} \quad \frac{1}{2n}||\tilde{\mathbf{X}} - \tilde{\mathbf{X}}W||_F^2 + \lambda||W||_1 \qquad \text{such that} \qquad h(W) = 0, \tag{6.9}$$

which is non-convex but can be solved approximately with second-order methods. [201] proposed to use an augmented Lagrangian method that we leverage here, with resulting solutions often very close to the true global minimum in practice and that scale to modern problem sizes with thousands of variables[2].

Choosing the regularization parameter λ with cross-validation is different than in the standard setting with no confounding. When using cross-validation, aiming for best prediction, the chosen λ would be typically too small since the best prediction would also try to capture the unwanted signal from $\mathbf{X}C$ in (6.4). To partially correct for this issue, cross-validation should be run on the adjusted data $\tilde{\mathbf{X}}$. This strategy should make the additional signal smaller and hence cross-validation aiming for best prediction is expected to perform reasonably well.

We call this causal discovery approach the Deconfounded Score method (**DECS**).

6.4 Experiments on synthetic data

Our goal in this section is to measure causal discovery performance in extensive experiments, and especially under violations of our assumptions.

Comparisons. We make comparisons with three causal discovery methods: the independence-based Fast Causal Inference (**FCI**) [163], **LGES** [58] that uses a decomposition of the covariance matrix followed by the GES algorithm, and **Notears** [201], the continuous optimization approach without adjustments (it is not specifically designed for unobserved confounding but serves to isolate the benefit / harm of adjusting for unobserved confounding with **DECS**). We note that the performance of non-convex optimization programs in the context of DAG learning, and the benefit of continuous-optimization formulations for DAG learning are well studied [201, 122] – both noting significant gains over independence-based methods.

Metric. Note however that not all algorithms have the same output, **FCI** outputs an equivalence class of MAGs, **LGES** outputs an equivalence class of DAGs, and **Notears** outputs a weighted adjacency matrix.

For consistent performance comparisons, we chose to consider the *skeleton* (i.e. all directionality omitted) of estimated graphs which is a common output across all algorithms. In a sense this treats existing algorithms favourably by regarding undirected or undetermined edges as true positives as long as the true graph has a directed edge in place of the undirected edge. (We give

[2]Recently, [122] found that enforcing $h(W) = 0$ may not be necessary to recover a DAG in practice, and argue for a soft constraint leading to much faster methods. One may extend the above in the same manner.

more details on algorithm and metric implementation in the Appendix). We report the AUC and SHD on estimated skeletons and both take into account false positives and false negatives.

We do make more detailed evaluations in the Appendix considering the error in weighted adjacency recovery $(W - \hat{W})^2$ (although comparisons there are limited to **Notears** which is the only baseline outputting weighted adjacency matrices).

6.4.1 Experimental set-up

In each experiment, we generated a p-dimensional random graph G from a Erdös–Rényi random graph model with p edges on average. Given G, we assigned uniformly random edge weights to obtain a weighted adjacency matrix $W \in \mathbb{R}^{p\times p}$. Given W, we sampled $X = WX + BH + E$ repeatedly from different noise models for $H \in \mathbb{R}^q$ and $E \in \mathbb{R}^p$, including Gaussian, Exponential and Gumbel distributions, and $B \in \mathbb{R}^{p\times q}$ with each entry independently sampled from $\mathcal{N}(0,1)$. We fix the number of observations $n = 100$ in all experiments.

Task. The task is to recover the skeleton defined by W (i.e. the matrix $\bar{W}$ such that $[\bar{W}]_{ij} = \mathbf{1}\{[W]_{ij} \neq 0\}$) given n independent samples from X, available as data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$. We consider performance comparisons along the spectrum of five parameters:

1. The data distribution family.
2. The dimensionality p of X.
3. The dimensionality q of H.
4. The noise scale σ which when small implies a more pronounced perturbation of unobserved confounding.
5. The denseness of B.

6.4.2 Results

(1) The data distribution family. Each column of Figure 6.2 refers to a different data distribution family. We can see that when the Gaussian assumption is satisfied DECS can significantly improve in performance with respect to other methods, especially for relatively high-dimensional graphs (top row). It is interesting however that relative performance does

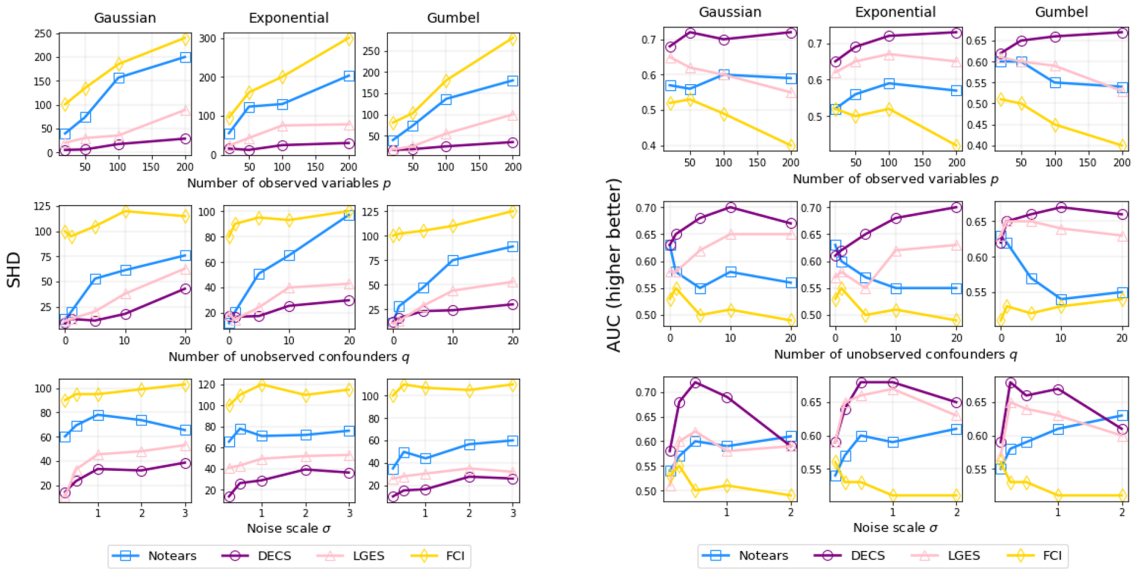


Figure 6.2: Performance on synthetic experiments. DECS is the proposed approach.

not vary with a change in distribution (Exponential of Gumbel) which suggests that DECS is relatively robust to the underlying noise model.

(2) Dimensionality of observed variables. In the top row of Figure 6.2 we show performance as a function of the dimensionality of the observables. Theoretically, DECS requires high-dimensional data and we see that outperformance is strongest in this regime (the number of samples here is 100) although DECS remains competitive otherwise.

(3 and 4) Dimensionality and strength of unobserved confounders. On the middle and bottom rows we vary the dimensionality of unobserved confounders q and strength of confounding (their variance σ) respectively. When $q = 0$ the system is fully observed. An interesting observation is that Notears and DECS perform similarly which suggests that there is nothing lost by adjusting for unobserved confounders even when not present. As we increase q and the strength of confounding DECS outperforms.

(5) Sparse unobserved confounders. In the Appendix we conduct an experiment to test the sensitivity of DECS with respect to the level of denseness on B. The advantage of DECS decreases in this case, though performance remains competitive.

6.5 Experiments on Genetic Data

The study of gene regulatory networks is one area in genomics with the potential to uncover the interactions of molecular regulators that govern the gene expression levels of messenger RNA and proteins: the building blocks of all cell function.

Problem. In this section, we are interested in the problem of recovering the underlying network from individual samples of gene expression. To validate performance on this task, we use a number of gene expression simulation programs that have been constructed based on the behaviour of known simple organisms, all publicly available in the `bnlearn` R package.

Data. We consider gene expression data from an E. coli microorganism [152] (**E. coli**), gene expression data describing starch metabolism of Arabidopsis thaliana [125] (**Starch**), data from a scale-free network, found to faithfully describe biological organisms [7] (**Scale-Free**), and protein expression level data from human immune system cells [148] (**Sachs**). All variables are fully observed in all of the above. We consider inducing unobserved confounding by explicitly removing a number root nodes in the network after sampling data, see Figure 6.3 for an example with the Starch network. Networks, omitted variables, and other details for all datasets can be found in the Appendix.

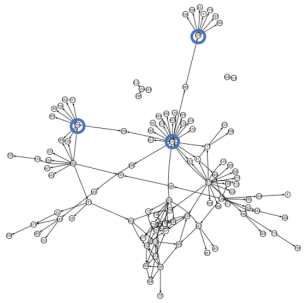


Figure 6.3: **Starch** network. Data from each of the blue nodes in the starch network is omitted thereby inducing spurious correlations among their children.

Results. Performance results are given in Table 6.1. AUC and SHD figures on all datasets show that DECS is competitive on all tasks. We make an additional comparison here considering true positive (TPR) and false discovery (FDR) rates at a threshold chosen for minimum SHD. This comparison is made to show the relatively good false discovery control of DECS even though formal guarantees were not established. On all metrics, and particularly with the AUC that considers performance along the whole threshold spectrum, DECS outperforms in most cases which demonstrates its applicability in realistic genetic data scenarios where knowledge on interactions between genes or gene products are typically not available without interventions.

Table 6.1: Mean performance and standard deviations over 10 random trials on Ecoli ($n = 100, p = 41$), Starch ($n = 100, p = 104$), Scale-Free ($n = 100, p = 200$) and Sachs ($n = 7466, p = 10$) data. Bold indicates best performance. FCI returns a complete graph in almost all cases and we have omitted it from these results.

		E. coli	**Starch**	**Scale-Free**	**Sachs**
TPR	Notears	0.39 ± 0.01	0.24 ± 0.01	0.18 ± 0.01	0.58 ± 0.02
	LGES	$\mathbf{0.57 \pm 0.05}$	$\mathbf{0.42 \pm 0.03}$	0.15 ± 0.01	$\mathbf{0.66 \pm 0.05}$
	DECS (ours)	0.34 ± 0.04	0.28 ± 0.05	$\mathbf{0.21 \pm 0.05}$	0.33 ± 0.05
FDR	Notears	0.59 ± 0.02	0.83 ± 0.05	$\mathbf{0.12 \pm 0.01}$	0.64 ± 0.05
	LGES	0.66 ± 0.03	0.58 ± 0.03	0.82 ± 0.05	0.55 ± 0.05
	DECS (ours)	$\mathbf{0.32 \pm 0.05}$	$\mathbf{0.50 \pm 0.06}$	0.20 ± 0.03	$\mathbf{0.20} \pm 0.04$
SHD	Notears	39.0 ± 2.25	192 ± 10.0	40.0 ± 5.25	12.5 ± 1.50
	LGES	51.0 ± 2.50	115 ± 7.00	23.0 ± 2.00	13.0 ± 1.50
	DECS (ours)	$\mathbf{26.0 \pm 2.00}$	$\mathbf{95.0 \pm 3.00}$	$\mathbf{14.0 \pm 1.25}$	$\mathbf{8.00 \pm 1.00}$
AUC	Notears	0.60 ± 0.02	0.58 ± 0.03	0.65 ± 0.03	$\mathbf{0.67 \pm 0.03}$
	LGES	0.62 ± 0.05	0.66 ± 0.06	0.59 ± 0.05	0.66 ± 0.04
	DECS (ours)	$\mathbf{0.65 \pm 0.03}$	$\mathbf{0.67 \pm 0.04}$	$\mathbf{0.70 \pm 0.05}$	0.65 ± 0.05

6.5.1 DECS for reproducible discovery

This section considers reproducibility of causal discovery across environments.

If two datasets differ in the distribution of unmeasured variation, correlations between observables vary, and we cannot expect estimates of conventional causal discovery algorithms to be reproducible. This is an important challenge because any two experiments most likely do differ due to changes in environment, data collection practices, among other unmeasured factors. The adjusted adjacency matrix from DECS, by definition removes sources of unmeasured variation from the otherwise biased estimate. We can expect the estimated adjacency matrix to be invariant in theory to changes in distribution of unobserved confounders, and therefore more reproducible and stable across different experiments.

Experiment design. To test this feature, we adopt the scale-free network and construct several datasets while varying the extent of unobserved confounding to simulate different environments. Specifically, we let $X = WX + BH + E$, where matrices W and B, and the distribution $E \sim \mathcal{N}_p(0, I)$ are fixed, while H is drawn from distributions $\mathcal{N}(0, \sigma)$ with varying σ (drawn at random in the interval $[0.25, 2]$). Different σ correspond to different environments. The problem is to test the agreement between recovered adjacency matrices W in different environments.

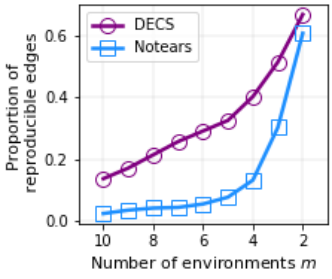


Figure 6.4: Reproducibility experiments. Higher values for larger number of environments indicate higher levels of reproducibility.

Results. We report the number of edges that reproduce across different environments in Figure 6.4. Each point on the plot gives the proportion of estimated edges that intersect in any m studies, $m = 1, \ldots, 10$. For instance, approximately 15% of estimated edges (across all 10 environments) intersect in all 10 environments for DECS whereas only 1% do for Notears[3]. This shows that adjusting for unobserved confounding improves the reproducibility of causal discovery.

6.6 Conclusions

This chapter develops a score-based causal discovery algorithm in the presence of dense unobserved confounding (unobserved variables with a widespread effect on observed ones). The argument considers properties of the spectrum of the data matrix that allows DAG learning (directed edges among observed variables) in the presence of dense confounding to be expressed as a continuous optimization problem. Solutions to this problem have guarantees on the true positive rate in the high-dimensional regime, the resulting score-based problem is much simpler to implement than independence-based alternatives and it outperforms empirically across a range of different experiments.

One may extend the proposed approach to model more general structural models. Specifically, structural models not constrained by a specific data distribution family or functional relationships between variables. One may consider as an extension optimization problems of the form,

$$\underset{f \in \mathcal{F}}{\arg\min} \quad \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(\tilde{\mathbf{x}}_i, f(\tilde{\mathbf{x}}_i)\right) + \rho_\lambda(f), \tag{6.10}$$

where $\mathcal{F}$ is a more general space of functions $f : \mathbb{R}^p \to \mathbb{R}^p$ that defines the causal structure in the data through its partial derivatives with respect to its arguments, $\mathcal{L} : \mathbb{R}^{p \times p} \to \mathbb{R}$ is a loss function (that may be chosen to model other data types, such as binary or count data) and $\rho_\lambda(f)$ is a regularization term that includes the acyclicity constraint. In this case, it takes a different form but may be computed for large classes of functions by considering norms on partial derivatives as in [202] and has already been shown to be successful for non-linear models in the fully observed setting.

[3]This experiment considers adjacency matrix recovery but we make additional comparisons on the basis of skeleton recovery with LGES in the Appendix.

There is scope as well for considering other adjustment frameworks that control for the influence of unobserved confounding. For instance, using different problem-dependent eigenvalue thresholds in the adjusted data matrix or by optimizing simultaneously for matrices W and B in the linear structural model with an l_1 and l_2 penalty respectively as [36] considered in the regression setting.

References

[1] Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

[2] Philip W Anderson. More is different. *Science*, 177(4047):393–396, 1972.

[3] Bryon Aragam, Arash Amini, and Qing Zhou. Globally optimal score-based learning of directed acyclic graphs in high-dimensions. In *Advances in Neural Information Processing Systems*, pages 4450–4462, 2019.

[4] Bryon Aragam, Arash A Amini, and Qing Zhou. Learning directed acyclic graphs with penalized neighbourhood regression. *arXiv preprint arXiv:1511.08963*, 2015.

[5] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[6] Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003.

[7] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[8] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108, 2012.

[9] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603, 2012.

[10] Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R Devon Hjelm, and Aaron Courville. Mine: mutual information neural estimation. In *International Conference on Machine Learning*, 2018.

[11] Alexis Bellot and Mihaela van der Schaar. Conditional independence testing using generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 2202–2211, 2019.

[12] Alexis Bellot and Mihaela van der Schaar. Kernel hypothesis testing with set-valued data, 2019.

[13] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[14] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

[15] Alessio Benavoli and Francesca Mangili. Gaussian processes for bayesian hypothesis tests on regression functions. In *Artificial intelligence and statistics*, pages 74–82, 2015.

[16] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[17] Daniel Bernstein, Basil Saeed, Chandler Squires, and Caroline Uhler. Ordering-based causal structure learning in the presence of latent variables. In *International Conference on Artificial Intelligence and Statistics*, pages 4098–4108. PMLR, 2020.

[18] Thomas B Berrett and Richard J Samworth. Nonparametric independence testing via mutual information. *Biometrika*, 106(3):547–566, 2019.

[19] Thomas B Berrett and Richard J Samworth. Usp: an independence test that improves on pearson's chi-squared and the *g*-test. *arXiv preprint arXiv:2101.10880*, 2021.

[20] Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test. *arXiv preprint arXiv:1807.05405*, 2018.

[21] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(Sep):2137–2155, 2009.

[22] Leslie G Biesecker. Hypothesis-generating research and predictive medicine. *Genome research*, 23(7):1051–1053, 2013.

[23] Xin Bing, Florentina Bunea, Yang Ning, Marten Wegkamp, et al. Adaptive estimation in structured factor models with applications to overlapping clustering. *Annals of Statistics*, 48(4):2055–2081, 2020.

[24] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv preprint arXiv:1711.07910*, 2017.

[25] J Martin Bland and Sally M Kerry. Weighted comparison of means. *Bmj*, 316(7125):129, 1998.

[26] Allan M Brandt. *The cigarette century: the rise, fall, and deadly persistence of the product that defined America*. Basic Books (AZ), 2007.

[27] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[28] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. Unintended consequences of machine learning in medicine. *Jama*, 318(6):517–518, 2017.

[29] T Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128, 2021.

[30] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

[31] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[32] Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. CRC press, 2006.

[33] Domagoj Ćevid, Peter Bühlmann, and Nicolai Meinshausen. Spectral deconfounding via perturbed sparse linear models. *arXiv preprint arXiv:1811.05352*, 2018.

[34] Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets, 1982.

[35] Anindya Chatterjee, Joydeep Ghosh, Baskar Ramdas, Raghuveer Singh Mali, Holly Martin, Michihiro Kobayashi, Sasidhar Vemula, Victor H Canela, Emily R Waskow, Valeria Visconte, et al. Regulation of stat5 by fak and pak1 in oncogenic flt3-and kit-driven leukemogenesis. *Cell reports*, 9(4):1333–1348, 2014.

[36] Victor Chernozhukov, Christian Hansen, Yuan Liao, et al. A lava attack on the recovery of sums of dense and sparse signals. *Annals of Statistics*, 45(1):39–76, 2017.

[37] David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *Journal of machine learning research*, 2(Feb):445–498, 2002.

[38] Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. In *Advances in neural information processing systems*, pages 406–414, 2010.

[39] Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015.

[40] Tom Claassen, Joris Mooij, and Tom Heskes. Learning sparse causal models is not np-hard. *arXiv preprint arXiv:1309.6824*, 2013.

[41] Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, 2014.

[42] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[43] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, 2008.

[44] Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010.

[45] Bruno De Finetti. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68, 1937.

[46] T De Wet, JH Venter, et al. Asymptotic distributions for quadratic forms with applications to tests of fit. *The Annals of Statistics*, 1(2):380–387, 1973.

[47] Persi Diaconis and David Freedman. Finite exchangeable sequences. *The Annals of Probability*, pages 745–764, 1980.

[48] Maurice Diesendruck, Guy W Cole, and Sinead Williamson. Directing generative networks with weighted maximum mean discrepancy. 2018.

[49] Peng Ding, Avi Feller, and Luke Miratrix. Randomization inference for treatment effect variation. *arXiv preprint arXiv:1412.5000*, 2014.

[50] Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pages 132–141, 2014.

[51] Shayan Doroudi, Philip S Thomas, and Emma Brunskill. Importance sampling for fair policy selection. *Grantee Submission*, 2017.

[52] Mathias Drton, Michael Eichler, and Thomas S Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(10), 2009.

[53] John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

[54] John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

[55] John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts. *Under review*, 2019.

[56] Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 75(4), 2013.

[57] Jianqing Fan, Han Liu, and Weichen Wang. Large covariance estimation through elliptical factor models. *Annals of statistics*, 46(4):1383, 2018.

[58] Benjamin Frot, Preetam Nandy, and Marloes H Maathuis. Robust causal structure learning with some hidden variables. *arXiv preprint arXiv:1708.01151*, 2017.

[59] Wayne A Fuller. *Measurement error models*, volume 305. John Wiley & Sons, 2009.

[60] Johann A Gagnon-Bartsch, Laurent Jacob, and Terence P Speed. Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California*, pages 1–112, 2013.

[61] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[62] Rui Gao, Liyan Xie, Yao Xie, and Huan Xu. Robust hypothesis testing using wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, pages 7902–7912, 2018.

[63] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570, 2012.

[64] AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. In *Advances in Neural Information Processing Systems*, pages 3011–3021, 2017.

[65] Jaime Roquero Gimenez and James Zou. Identifying invariant factors across multiple environments with kl regression. *arXiv preprint arXiv:2002.08341*, 2020.

[66] Jelle J Goeman, Sara A Van De Geer, Floor De Kort, and Hans C Van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[67] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[68] Maria Elizabeth Grabe, KD Trager, Melissa Lear, and Jennifer Rauch. Gender in crime news: A case study test of the chivalry hypothesis. *Mass Communication & Society*, 9(2):137–163, 2006.

[69] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007.

[70] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[71] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[72] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pages 673–681, 2009.

[73] Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592, 2008.

[74] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 2009.

[75] Gökhan Gül and Abdelhak M Zoubir. Robust hypothesis testing with $\backslash\alpha$-divergence. *IEEE Transactions on Signal Processing*, 64(18):4737–4750, 2016.

[76] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[77] James J Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

[78] Harold Hotelling. The generalization of student's ratio. In *Breakthroughs in statistics*, pages 54–65. Springer, 1992.

[79] Harold Hotelling et al. A generalized t test and measure of multivariate dispersion. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*. The Regents of the University of California, 1951.

[80] Janet Shibley Hyde. The gender similarities hypothesis. *American psychologist*, 60(6):581, 2005.

[81] Tommi Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Advances in neural information processing systems*, pages 487–493, 1999.

[82] Dominik Janzing and Bernhard Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1), 2018.

[83] Dominik Janzing and Bernhard Schölkopf. Detecting non-causal artifacts in multivariate linear regression models. *arXiv preprint arXiv:1803.00810*, 2018.

[84] Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *Journal of Machine Learning Research*, 5(Jul):819–844, 2004.

[85] Wittawat Jitkrittum, Zoltén Szabó, and Arthur Gretton. An adaptive test of independence with analytic kernel embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1742–1751. JMLR. org, 2017.

[86] Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pages 262–271, 2017.

[87] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

[88] Philips George John, Deepak Vijaykeerthy, and Diptikalyan Saha. Verifying individual fairness in machine learning models. In *Conference on Uncertainty in Artificial Intelligence*, pages 749–758. PMLR, 2020.

[89] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Knockoffgan: Generating knockoffs for feature selection using generative adversarial networks. In *ICLR*, 2019.

[90] Eitan Kerem, Joseph Reisman, Mary Corey, Gerard J Canny, and Henry Levison. Prediction of mortality in patients with cystic fibrosis. *New England Journal of Medicine*, 326(18):1187–1191, 1992.

[91] Eitan Kerem, Laura Viviani, Anna Zolin, Stephanie MacNeill, Elpis Hatziagorou, Helmut Ellemunter, Pavel Drevinek, Vincent Gulmans, Uros Krivec, and Hanne Olesen. Factors associated with fev1 decline in cystic fibrosis: analysis of the ecfs patient registry. *European Respiratory Journal*, 43(1):125–133, 2014.

[92] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[93] Amit V Khera and Sekar Kathiresan. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nature reviews Genetics*, 18(6):331, 2017.

[94] Niki Kilbertus, Philip J Ball, Matt J Kusner, Adrian Weller, and Ricardo Silva. The sensitivity of counterfactual fairness to unmeasured confounding. In *Uncertainty in Artificial Intelligence*, pages 616–626. PMLR, 2020.

[95] Risi Kondor and Tony Jebara. A kernel between sets of vectors. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 361–368, 2003.

[96] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

[97] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

[98] John Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6(Jan):129–163, 2005.

[99] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

[100] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[101] Ho Chung Law, Christopher Yau, and Dino Sejdinovic. Testing and learning on distributions with symmetric noise invariance. In *Advances in Neural Information Processing Systems*, pages 1343–1353, 2017.

[102] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

[103] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[104] Giovanni Leoni. *A first course in Sobolev spaces*. American Mathematical Soc., 2017.

[105] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[106] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[107] Xihong Lin. Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326, 1997.

[108] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and DJ Sutherland. Learning deep kernels for non-parametric two-sample tests. *arXiv preprint arXiv:2002.09116*, 2020.

[109] James R Lloyd and Zoubin Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, pages 829–837, 2015.

[110] Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

[111] Miles Lopes, Laurent Jacob, and Martin J Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, pages 1206–1214, 2011.

[112] David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. The randomized dependence coefficient. In *Advances in neural information processing systems*, pages 1–9, 2013.

[113] David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pages 1452–1461, 2015.

[114] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2016.

[115] MAGGIC. The survival of patients with heart failure with preserved or reduced left ventricular ejection fraction: an individual patient data meta-analysis. *European heart journal*, 33(14):1750–1757, 2012.

[116] Subha Maity, Yuekai Sun, and Moulinath Banerjee. Minimax optimal approaches to the label shift problem. *arXiv preprint arXiv:2003.10443*, 2020.

[117] Nicolai Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10. IEEE, 2018.

[118] Nicolai Meinshausen, Peter Bühlmann, et al. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

[119] Pedro J Moreno, Purdy P Ho, and Nuno Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Advances in neural information processing systems*, pages 1385–1392, 2004.

[120] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *Advances in neural information processing systems*, pages 10–18, 2012.

[121] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *arXiv preprint arXiv:1605.09522*, 2016.

[122] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *arXiv preprint arXiv:2006.10201*, 2020.

[123] E Noether. Invariante variationsprobleme. nachr. vd ges. d. wiss. zu göttingen (1918) 235; e. noether e ma tavel. *Transport Theor. Stat. Phys*, 1:183, 1971.

[124] Christopher Nowzohour, Marloes Maathuis, and Peter Bühlmann. Structure learning with bow-free acyclic path diagrams. *stat*, 1050:7, 2015.

[125] Rainer Opgen-Rhein and Korbinian Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC systems biology*, 1(1):1–10, 2007.

[126] Victor M Panaretos, David Kraus, and John H Maddocks. Second-order comparison of gaussian random functions and the geometry of dna minicircles. *Journal of the American Statistical Association*, 105(490):670–682, 2010.

[127] Judea Pearl. Why there is no statistical test for confounding, why many think there is, and why they are almost right. 1998.

[128] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

[129] Judea Pearl. *Causality*. Cambridge university press, 2009.

[130] Judea Pearl. On a class of bias-amplifying variables that endanger effect estimates. *arXiv preprint arXiv:1203.3503*, 2012.

[131] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[132] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[133] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[134] Todd D Prickett, Brad Zerlanko, Jared J Gartner, Stephen CJ Parker, Ken Dutton-Regester, Jimmy C Lin, Jamie K Teer, Xiaomu Wei, Jiji Jiang, Guo Chen, et al. Somatic mutations in map3k5 attenuate its proapoptotic function in melanoma through increased binding to thioredoxin. *Journal of Investigative Dermatology*, 134(2):452–460, 2014.

[135] Robert N Proctor and Robert Proctor. *Golden holocaust: origins of the cigarette catastrophe and the case for abolition*. Univ of California Press, 2011.

[136] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[137] Anant Raj, Ho Chung Leon Law, Dino Sejdinovic, and Mijung Park. A differentially private kernel two-sample test. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 697–724. Springer, 2019.

[138] Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry A Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI*, pages 3571–3577, 2015.

[139] Thomas Richardson, Peter Spirtes, et al. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

[140] Paul R Rider. On the distribution of the correlation coefficient in small samples. *Biometrika*, pages 382–403, 1932.

[141] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

[142] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[143] Dominik Rothenhäusler, Peter Bühlmann, Nicolai Meinshausen, et al. Causal dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. *The Annals of Statistics*, 47(3):1688–1722, 2019.

[144] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

[145] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[146] Walter Rudin. *Fourier analysis on groups*, volume 121967. Wiley Online Library, 1962.

[147] Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *AISTATS*, 2018.

[148] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[149] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[150] Betul Erdogdu Sakar, M Erdem Isenkul, C Okan Sakar, Ahmet Sertbas, Fikret Gurgen, Sakir Delil, Hulya Apaydin, and Olcay Kursun. Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4):828–834, 2013.

[151] Robert Schall. Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–727, 1991.

[152] W Schmidt-Heck, R Guthke, S Toepfer, H Reischer, K Duerrschmid, and K Bayer. Reverse engineering of the stress response during expression of a recombinant protein. In *Proceedings of the EUNITE symposium*, pages 10–12, 2004.

[153] Marco Scutari, Catharina Elisabeth Graafland, and José Manuel Gutiérrez. Who learns better bayesian network structures: Constraint-based, score-based or hybrid algorithms? In *International Conference on Probabilistic Graphical Models*, pages 416–427, 2018.

[154] Dino Sejdinovic, Arthur Gretton, and Wicher Bergsma. A kernel test for three-variable interactions. *arXiv preprint arXiv:1306.2281*, 2013.

[155] Rajat Sen, Karthikeyan Shanmugam, Himanshu Asnani, Arman Rahimzamani, and Sreeram Kannan. Mimic and classify: A meta-algorithm for conditional independence testing. *arXiv preprint arXiv:1806.09708*, 2018.

[156] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Advances in Neural Information Processing Systems*, pages 2951–2961, 2017.

[157] Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

[158] Rajen D Shah, Benjamin Frot, Gian-Andrea Thanei, and Nicolai Meinshausen. Right singular vector projection graphs: fast high dimensional covariance matrix estimation under latent confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2020.

[159] Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *arXiv preprint arXiv:1804.07203*, 2018.

[160] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

[161] Connie P Shapiro, Lawrence Hubert, et al. Asymptotic normality of permutation statistics derived from weighted sums of bivariate functions. *The Annals of Statistics*, 7(4):788–794, 1979.

[162] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[163] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[164] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, Gert RG Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[165] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

[166] Thomas Stocker. *Introduction to climate modelling*. Springer Science & Business Media, 2011.

[167] Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *arXiv preprint arXiv:1702.03877*, 2017.

[168] Masashi Sugiyama et al. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, pages 985–1005, 2007.

[169] Dougal J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*, 2016.

[170] Zoltán Szabó, Arthur Gretton, Barnabás Póczos, and Bharath Sriperumbudur. Two-stage sampled learning theory on distributions. In *Artificial Intelligence and Statistics*, pages 948–957, 2015.

[171] Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

[172] Barbara G Tabachnick, Linda S Fidell, and Jodie B Ullman. *Using multivariate statistics*, volume 7. Pearson Boston, MA, 2019.

[173] Wesley Tansey, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M Blei. The holdout randomization test: Principled and easy black box feature selection. *arXiv preprint arXiv:1811.00645*, 2018.

[174] David Taylor-Robinson, Margaret Whitehead, Finn Diderichsen, Hanne Vebert Olesen, Tania Pressler, Rosalind L Smyth, and Peter Diggle. Understanding the natural progression in% fev1 decline in patients with cystic fibrosis: a longitudinal study. *Thorax*, 67(10):860–866, 2012.

[175] Joy A Thomas and TM Cover. Elements of information theory. *John Wiley & Sons, Inc., New York. Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, MPH (2009),"Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," Journal of the Royal Society Interface*, 6:187–202, 1991.

[176] Sofia Triantafillou and Ioannis Tsamardinos. Score-based vs constraint-based causal learning in the presence of confounders. In *CFA@ UAI*, pages 59–67, 2016.

[177] Amal Rannen Triki, Maxim Berman, and Matthew B Blaschko. Function norms and regularization in deep networks. *arXiv preprint arXiv:1710.06703*, 2017.

[178] Konstantinos Tsirlis, Vincenzo Lagani, Sofia Triantafillou, and Ioannis Tsamardinos. On scoring maximal ancestral graphs with the max–min hill climbing algorithm. *International Journal of Approximate Reasoning*, 102:74–85, 2018.

[179] Aad W van der Vaart and Jon A Wellner. The delta-method. In *Weak Convergence and Empirical Processes*, pages 372–400. Springer, 1996.

[180] Subhashini Venugopalan, Arunachalam Narayanaswamy, Samuel Yang, Anton Gerashcenko, Scott Lipnick, Nina Makhortova, James Hawrot, Christine Marques, Joao Pereira, Michael Brenner, et al. It's easy to fool yourself: Case studies on identifying bias and confounding in bio-medical datasets. *arXiv preprint arXiv:1912.07661*, 2019.

[181] Lois M Verbrugge. Gender and health: an update on hypotheses and evidence. *Journal of health and social behavior*, pages 156–182, 1985.

[182] Steve Verrill and Richard A Johnson. Asymptotic distributions for quadratic forms with applications to censored data tests of fit. *Communications in Statistics-Theory and Methods*, 17(12):4011–4024, 1988.

[183] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[184] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, pages 5334–5344, 2018.

[185] Jeffrey S Wagener, Michael J Williams, Stefanie J Millar, Wayne J Morgan, David J Pasta, and Michael W Konstan. Pulmonary exacerbations and acute declines in lung function in patients with cystic fibrosis. *Journal of Cystic Fibrosis*, 17(4):496–502, 2018.

[186] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[187] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[188] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[189] David Wozabal. A framework for optimization under ambiguity. *Annals of Operations Research*, 193(1):21–47, 2012.

[190] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2272–2281, 2017.

[191] Hongliang Yan, Zhetao Li, Qilong Wang, Peihua Li, Yong Xu, and Wangmeng Zuo. Weighted and class-specific maximum mean discrepancy for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 22(9):2420–2433, 2019.

[192] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in neural information processing systems*, pages 3391–3401, 2017.

[193] Jerrold H Zar. Significance testing of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, 67(339):578–580, 1972.

[194] Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in neural information processing systems*, pages 755–763, 2013.

[195] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric K Oermann. Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint arXiv:1807.00431*, 2018.

[196] Jin-Ting Zhang. Statistical inferences for linear models with functional responses. *Statistica Sinica*, pages 1431–1451, 2011.

[197] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *UAI*, 2012.

[198] Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018.

[199] Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. Learning overlapping representations for the estimation of individualized treatment effects. *arXiv preprint arXiv:2001.04754*, 2020.

[200] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. In *ICLR*, 2017.

[201] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pages 9472–9483, 2018.

[202] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.

[203] Zhihong Zhu, Zhili Zheng, Futao Zhang, Yang Wu, Maciej Trzaskowski, Robert Maier, Matthew R Robinson, John J McGrath, Peter M Visscher, Naomi R Wray, et al. Causal associations between risk factors and common diseases inferred from gwas summary data. *Nature communications*, 9(1):224, 2018.

[204] Indre Zliobaite. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*, 2015.

Appendix A

Appendix to Chapter 2

This appendix provides additional material to supplement the main body of this paper. It is outlined as follows:

- Section A.1 provides the proofs for all statements made in the main body of this paper.
  - Section A.1.1 gives the proof of the consistency of the RMMD.
  - Section A.1.2 gives the proof of the consistency of the RHSIC.
- Section A.2 gives details on the approximations used to deal with irregular set sizes and high-dimensional data.
- Section A.3 gives details on the implementation of baseline tests.

A.1 Proofs

A.1.1 Asymptotic distribution of $\widehat{\text{RMMD}}^2$

Our proof strategy consists of demonstrating convergence in probability of each inner product $K(\hat{\mu}_{\mathbb{P}}, \hat{\mu}_{\mathbb{Q}})$ to its population counterpart $K(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}})$, and take also into account approximations to the embeddings themselves we might make such as with Fourier features. Given convergence in probability (at a fast enough rate), the equivalence of their asymptotic distributions then follows by convergence results of random variables.

Background

All results in this section consider the asymptotic regime of increasing sample size N and increasing set size n_i for each i. We therefore make abstraction for notational simplicity of our weighting mechanism, assumed fixed and each weight identical across sets asymptotically which is equivalent to reverting to the equal weight scenario for our asymptotic results.

We start by recalling some definitions. The empirical statistic of the RMMD is given by,

$$\widehat{\mathrm{RMMD}}^2 := \frac{1}{N^2}\sum_{i,j=1}^{N} K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j}) + \frac{1}{M^2}\sum_{i,j=1}^{M} K(\hat{\mu}_{\mathbb{Q}_i}, \hat{\mu}_{\mathbb{Q}_j}) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{Q}_j}), \quad \text{(A.1)}$$

while the MMD with population mean embeddings is given by,

$$\widehat{\mathrm{MMD}}^2 := \frac{1}{N^2}\sum_{i,j=1}^{N} K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}) + \frac{1}{M^2}\sum_{i,j=1}^{M} K(\mu_{\mathbb{Q}_i}, \mu_{\mathbb{Q}_j}) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{Q}_j}). \quad \text{(A.2)}$$

We assume without loss of generality that $N = M$ for notational simplicity.

Let us recall also the asymptotic distributions under the null and alternative of the $\widehat{\mathrm{MMD}}^2$ given by [70].

Theorem [70]. Assume that K has finite second moments. Then, the following statements hold.

1. Under $\mathcal{H}_0$, $N\widehat{\mathrm{MMD}}^2 \xrightarrow{d} \sum_{l=1}^{\infty} \lambda_l (z_l^2 - 2)$. z_l is a sequence of Gaussian random variables and λ_l are the eigenvalues solution to a certain eigenvalue problem.
2. Under $\mathcal{H}_1$, $N^{1/2}\left(\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}^2\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\mathcal{H}_1}\right)$.

Please find the details of the eigenvalues and asymptotic variance in [70].

Now note that,

$$N\widehat{\mathrm{RMMD}}^2 = N\widehat{\mathrm{MMD}}^2 + (N\widehat{\mathrm{RMMD}}^2 - N\widehat{\mathrm{MMD}}^2)$$
$$\sqrt{N}\widehat{\mathrm{RMMD}}^2 = \sqrt{N}\widehat{\mathrm{MMD}}^2 + (\sqrt{N}\widehat{\mathrm{RMMD}}^2 - \sqrt{N}\widehat{\mathrm{MMD}}^2).$$

The first term relates to the asymptotic distribution of the RMMD under the null and the second term relates to the distribution of the RMMD under the alternative hypothesis.

We are interested in bounding the contribution of the second term in each case under the null and alternative hypotheses asymptotically. The absolute differences we are interested in bounding then under the null hypothesis given by,

$$\left|N\widehat{\mathrm{RMMD}}^2 - N\widehat{\mathrm{MMD}}^2\right| \leq \frac{1}{N}\sum_{i,j=1}^{N}\left|K(\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j}) - K(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j})\right| + \frac{1}{N}\sum_{i,j=1}^{N}\left|K(\mu_{\mathbb{Q}_i},\mu_{\mathbb{Q}_j}) - K(\hat{\mu}_{\mathbb{Q}_i},\hat{\mu}_{\mathbb{Q}_j})\right| - \frac{2}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|K(\mu_{\mathbb{P}_i},\mu_{\mathbb{Q}_j}) - K(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{Q}_j})\right|, \tag{A.3}$$

and under the alternative hypothesis,

$$\begin{aligned}\left|\sqrt{N}\widehat{\mathrm{RMMD}}^2 - \sqrt{N}\widehat{\mathrm{MMD}}^2\right| \leq &\frac{1}{N\sqrt{N}}\sum_{i,j=1}^{N}\left|K(\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j}) - K(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j})\right| \\ &+ \frac{1}{N\sqrt{N}}\sum_{i,j=1}^{N}\left|K(\mu_{\mathbb{Q}_i},\mu_{\mathbb{Q}_j}) - K(\hat{\mu}_{\mathbb{Q}_i},\hat{\mu}_{\mathbb{Q}_j})\right| \\ &- \frac{2}{N\sqrt{N}}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|K(\mu_{\mathbb{P}_i},\mu_{\mathbb{Q}_j}) - K(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{Q}_j})\right|.\end{aligned} \tag{A.4}$$

In both cases it suffices to show that inner products between population mean embeddings and empirical counterparts converge in probability at a rate fast enough such that a union bound over all terms in the summation scaled by $1/N$ and $1/(N\sqrt{N})$ converges to 0. We note here that we are considering two asymptotic regimes, once in the size of each set n_i that is relevant in the convergence of $K(\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j})$ to $K(\mu_{\mathbb{P}_i},\mu_{\mathbb{P}_j})$ and one in N which is the number of sets. Each may vary independently, and here we will assumed that the rate of growth of n_i is sufficient to ensure the weighted sums converge as $N \to \infty$.

Results

We will traverse the convergence of empirical kernels to their population counterparts in two steps, first using results that show the convergence of empirical mean embeddings to their population counterparts (Lemma 1) and second, using a Lipschitz condition to extend this to inner products between mean embeddings (Lemma 2).

For this we will assume K to be a real-valued, shift invariant ($K(x,x') = K(x-x',0)$), and L_K-Lipschitz kernel,

$$|K(x,0) - K(x',0)| \leq L_K|x-x'|, \tag{A.5}$$

also satisfying the boundedness condition $|K(x,x')| < 1$ for all $x, x' \in \mathcal{X}$.

The following two Lemmas demonstrate our claim.

Lemma 1 (Bound on the empirical mean embedding [113]) *Let the kernel K satisfy the assumptions above. Then we have,*

$$|\mu_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_i}|_{\mathcal{H}_K} \leq 2\sqrt{\frac{\mathbb{E}_{x\sim\mathbb{P}_i}K(x,x)}{n_i}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n_i}}, \tag{A.6}$$

with probability at least $1-\delta$ over the randomness in the empirical sample from $\mathbb{P}_i$. n_i is the number of samples from $\mathbb{P}_i$.

Lemma 2 (Bound on kernels computed on empirical mean embeddings) *Let K be defined as above. The it holds that,*

$$|K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}) - K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j})| \leq L_K \left(4\sqrt{\frac{1}{\eta}} + 2\sqrt{\frac{2\log\frac{1}{\delta}}{\eta}}\right), \tag{A.7}$$

with probability at least $1-\delta$. As $\eta := \min(n_i, n_j) \to \infty$ we get that $K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j})$ converges in probability to $K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j})$.

Proof. The proof is based on the Lipschitz condition and the error bound on empirical mean embeddings with respect to their population counterparts.

$$\begin{aligned}
|K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}) - K(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j})| &= \left|K(\mu_{\mathbb{P}_i} - \mu_{\mathbb{P}_j}, 0) - K(\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_j}, 0)\right| && \text{(A.8)}\\
&\leq L_K \left|\mu_{\mathbb{P}_i} - \mu_{\mathbb{P}_j} - (\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_j})\right| && \text{(A.9)}\\
&\leq L_K \left|\mu_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_i}\right| + L_K \left|\mu_{\mathbb{P}_j} - \hat{\mu}_{\mathbb{P}_j}\right| && \text{(A.10)}\\
&\leq L_K \left(2\sqrt{\frac{\mathbb{E}_{x\sim\mathbb{P}_i}K(x,x)}{n_i}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n_i}} + 2\sqrt{\frac{\mathbb{E}_{x\sim\mathbb{P}_j}K(x,x)}{n_j}} + \sqrt{\frac{2\log\frac{1}{\delta}}{n_j}}\right) && \text{(A.11)}\\
&\leq L_K \left(4\sqrt{\frac{1}{\eta}} + 2\sqrt{\frac{2\log\frac{1}{\delta}}{\eta}}\right), && \text{(A.12)}
\end{aligned}$$

where $\eta := \min(n_i, n_j)$ and we have use the boundedness condition on K, $\mathbb{E}_{x\sim\mathbb{P}_i}K(x,x) \leq 1$.

The for a rate of of increase of n_i fast enough in comparison to n, each term in equations (A.3) and (A.4) converges to zero which implies that the asymptotic distributions of $N\widehat{\text{RMMD}}^2$, $\sqrt{N}\widehat{\text{RMMD}}^2$ and $\sqrt{N}\widehat{\text{MMD}}^2$, $N\widehat{\text{MMD}}^2$, coincide respectively.

Extension to approximations using random Fourier features

For completeness, in addition to considering convergence in distribution using empirical embeddings, we extend our analysis to include Fourier feature approximations in the empirical embeddings themselves and their asymptotic behaviour. To do so notice that we may write,

$$|k(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_i,m}, \hat{\mu}_{\mathbb{P}_j,m})| \leq \\ \left|k(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j})\right| + \left|k(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_i,m}, \hat{\mu}_{\mathbb{P}_j,m})\right|, \tag{A.13}$$

by the triangle inequality.

The following two lemmas are similar to the first two above but instead related the empirical mean embedding $\hat{\mu}_{\mathbb{P}_i}$ with its random Fourier feature approximation $\hat{\mu}_{\mathbb{P}_i,m}$.

Lemma 3 (Bound on the randomized empirical mean embedding [113]) *Let k be defined as above. For a fixed sample of size n_i from a probability distribution $\mathbb{P}_i$ on $\mathbb{R}^d$ and any $\delta > 0$, we have,*

$$|\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_i,m}|_{L^2(\mathbb{P})} \leq \frac{2}{\sqrt{m}}\left(1 + \sqrt{2\log n_i/\delta}\right), \tag{A.14}$$

with probability larger than $1-\delta$ over the randomness of the samples $(\omega_i, b_i)_{i=1}^m$.

Lemma 4 (Bound on kernels computed on approximated empirical mean embeddings) *Let k be defined as above. Then for any $\varepsilon > 0$ it holds that,*

$$\left|k(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_i,m}, \hat{\mu}_{\mathbb{P}_j,m})\right| \leq \frac{2L_k}{\sqrt{m}}\left(2 + 2\sqrt{2\log(\eta/\delta)}\right), \tag{A.15}$$

m is the number of random features, n_i and n_j are the number of observations in time series X_i and X_j respectively, and $\eta := \min(n_i, n_j)$. If further we assume that $\min(n_i, n_j)\exp\{-m\} \to 0$ as $n_i, n_j, m \to \infty$, then $k(\hat{\mu}_{\mathbb{P}_i,m}, \hat{\mu}_{\mathbb{P}_j,m})$ converges in probability to $k(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j})$.

Proof. The proof strategy is similar to Lemma 3, but for with a different bound on the difference between mean embeddings. We proceed as follows,

$$\begin{aligned}
\left|k(\hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j}) - k(\hat{\mu}_{\mathbb{P}_{i,m}}, \hat{\mu}_{\mathbb{P}_{j,m}})\right| &= \left|k(\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_j}, 0) - k(\hat{\mu}_{\mathbb{P}_{i,m}} - \hat{\mu}_{\mathbb{P}_{j,m}}, 0)\right| && \text{(A.16)}\\
&\leq L_k \left|\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_j} - (\hat{\mu}_{\mathbb{P}_{i,m}} - \hat{\mu}_{\mathbb{P}_{j,m}})\right| && \text{(A.17)}\\
&\leq L_k \left|\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_{i,m}}\right| + L_k \left|\hat{\mu}_{\mathbb{P}_j} - \hat{\mu}_{\mathbb{P}_{j,m}}\right| && \text{(A.18)}\\
&\leq \frac{2L_k}{\sqrt{m}} \left(2 + \sqrt{2\log(n_i/\delta)} + \sqrt{2\log(n_j/\delta)}\right) && \text{(A.19)}\\
&\leq \frac{2L_k}{\sqrt{m}} \left(2 + 2\sqrt{2\log(\eta/\delta)}\right), && \text{(A.20)}
\end{aligned}$$

where we have written $\eta := \min(n_i, n_j)$ and the inequalities hold with probability at least $(1-\delta)$ over the randomness of the samples $(\omega_i, b_i)_{i=1}^m$.

A.1.2 Asymptotic distribution of $\widehat{\text{RHSIC}}$

The asymptotic distribution of the RHSIC follows a very similar procedure since it can similarly bee decomposed in sums of kernels.

Proof. The $\widehat{\text{RHSIC}}$ may be written as a sum of V-statistics as follows [73],

$$\widehat{\text{RHSIC}} = \frac{1}{N^2}\sum_{i,j}^{N} \hat{K}_{ij}\hat{L}_{ij} + \frac{1}{N^4}\sum_{i,j,q,r}^{N} \hat{K}_{ij}\hat{L}_{qr} - \frac{2}{N^3}\sum_{i,j,q}^{N} \hat{K}_{ij}\hat{L}_{iq}, \tag{A.21}$$

where to avoid cluttering the notation we have written $\hat{K}_{ij} := K(\hat{\mu}_{\mathbb{P}_{i,m}}, \hat{\mu}_{\mathbb{P}_{j,m}})$ and $\hat{L}_{ij} := L(\mu_{Y_i,m}, \mu_{Y_j,m})$. Sums with two summation indices refer to double sums of all pairs of numbers drawn with replacement from $\{1, ..., N\}$, and similarly for three and four summation indices [73]. Similarly to the two sample problem, equality in asymptotic distribution may be shown by considering the absolute differences in the product of population and empirical kernels. That is, we are interested in bounding the following,

$$|\hat{K}_{ij}\hat{L}_{qr} - K_{ij}L_{qr}|, \tag{A.22}$$

for any quadruple of indices i, j, q, r.

Assuming as above that kernels K and L are Lipschitz functions it follows that their product is also Lipschitz,

$$\begin{aligned}
&|K(x,0)L(y,0)-K(x',0)L(y',0)| \\
&\qquad \leq |(K(x,0)-K(x',0))L(y,0)+(L(y,0)-L(y',0))K(x',0)| \\
&\qquad \leq |K(x,0)-K(x',0)|\cdot||L(y,0)||_{\mathcal{H}_L}+|L(y,0)-L(y',0)|\cdot||K(x',0)||_{\mathcal{H}_K} \\
&\qquad \leq L_K|x-x'|+L_L|y-y'|.
\end{aligned}$$

The same arguments and lemmas used in the two-sample case apply which proves the equivalence in asymptotic distributions of the $\widehat{\text{RHSIC}}$ and $\widehat{\text{HSIC}}$.

A.2 Approximations for high power

A.2.1 Kernel hyperparameters

For the two sample problem, let N be the number of samples in both groups, which simplifies the formulation of the asymptotic power of the $\widehat{\text{RMMD}}^2$. The following procedure mirrors [169].

Proposition 3 ***(Approximate power of $\widehat{RMMD}^2$).*** *Under $\mathcal{H}_1$, for large N and fixed r, the test power $Pr(N\widehat{\text{RMMD}}^2 > r) \approx 1-\Phi(\frac{r}{\sqrt{N}\sigma_{\text{RMMD}}}-\sqrt{N}\frac{\text{RMMD}^2}{\sigma_{\text{RMMD}}})$ where Φ denotes the cumulative distribution function of the standard normal distribution, σ^2_{RMMD} is the asymptotic variance under $\mathcal{H}_1$ for the $\widehat{RMMD}^2$.*

Consider the terms inside the *cdf* of the normal. Observe that the first term $\frac{r}{\sqrt{N}\sigma_{\text{RMMD}}} = \mathcal{O}(N^{-1/2})$ goes to 0 as $N \to \infty$, while the second term, $\sqrt{N}\frac{\text{RMMD}^2}{\sigma_{\text{RMMD}}} = \mathcal{O}(N^{1/2})$, dominates the first one for large N. As an approximation, for sufficiently large N, the parameters that maximize the test power are given by $\theta^* = \text{argmax}_\theta \ Pr(N\widehat{\text{RMMD}}^2 > r) \approx \frac{\text{RMMD}^2}{\sigma_{\text{RMMD}}}$. In our case θ includes the bandwidth parameter used to compute the mean embeddings and the bandwidth parameter used to compute the test statistic. The empirical estimate of the variance $\hat{\sigma}_{\text{RMMD}}$ that appears in our objective is approximated up to second order terms, as in [169]. Similar derivations hold for the power optimization of the HSIC with the exception that the definition of the HSIC requires optimization of two kernels, one for each set in our paired samples: K and L.

Note that since RMMD and σ_{RMMD} are unknown, to maintain the validity of the hypothesis test we divide the sample into a training set, used to estimate the ratio with $\frac{\widehat{\text{RMMD}}^2}{\hat{\sigma}_{\text{RMMD}}}$ and choose the kernel parameters, and a testing set used to perform the final hypothesis test with the learned kernels.

An analogous result holds for the approximate power of $\widehat{\text{RHSIC}}$.

A.2.2 Weighting scheme

Under the alternative hypothesis, the asymptotic variance of the proposed test statistics is well defined and given by asymptotic theory of V-Statistics (up to scaling) equal to $\text{Var}(\mathbb{E}K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}))$, see e.g. Theorem 5.5.1 [157]. To specify the set of weights that maximize power we may use the same reasoning to the section above and minimize the asymptotic variance.

With finite samples to approximate the mean embedding, assuming that all randomness comes from the number of samples available to estimate mean embeddings, its variance is proportional to $1/n_i$. The delta method (see e.g. [179]) may be applied on the bivariate sample $(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j})$ with the function K to conclude that the variance of each $K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j})$ is proportional to $1/(n_i \cdot n_j)$. Now, with a finite number of sets, or in other words a finite number of distributions, we approximate the expectation $\mathbb{E}K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j})$ with averages. Assuming that the covariance between any pair $K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j})$ and $K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_k})$ for any i, j, k does not vary by changing indices, that is, is fixed, weighting each term $K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j})$ with the inverse of its variance gives the lowest attainable variance $\text{Var}(\mathbb{E}K(\mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j}))$ in finite samples.

A.3 Additional details on experiments and implementation

A.3.1 Details on the data generation mechanisms

The inverse gamma distribution has appeared parameterized by one and two parameters. We choose the one-parameter distributions with density,

$$f(x;\mu) = \frac{x^{-\mu-1}}{\Gamma(\mu)} \exp(-1/x), \tag{A.23}$$

where $x \geq 0$, $\mu > 0$ and Γ is the gamma function.

A.3.2 RMMD and RHSIC

We create empirical kernel mean embeddings by concatenating data along each dimension. Each embedding has random features sampled to approximate a Gaussian kernel with length scale parameter σ^2. σ^2 is estimated by cross-validation on a grid of parameter values around the median of squared pairwise distances of the stacked data. In practice, we set the number of random features to $m = 50$ (larger amounts of random features show no significant performance improvements). The parameters of the kernel used for testing are similarly optimized via cross-validation by defining a grid of parameter values around the median of squared pairwise distances of computed random features. In summary, for each random feature length-scale we test with a number of test length-scales and choose the pair of parameters with best performance according to our power criterion. A summary of these tests' implementation is as follows.

1. For each observed set $\{x_{i,j}\}_{j=1}^{n_i} \sim \mathbb{P}_i$, compute its approximated mean embedding using a Fourier basis, with elements in the span of $(\cos(\langle \omega_j, x\rangle + b_j))_{j=1}^m$,

$$\hat{\mu}_{\mathbb{P}_i,m} = \frac{1}{n_i} \sum_{x \in \{x_{ij}\}_{j=1}^{n_i}} (\cos(\langle w_j, x\rangle + b_j))_{j=1}^m \in \mathbb{R}^m.$$

2. Compute weights that describe the confidence we have in each of the above approximations, $w_{\mathbb{P}_i} := n_i / \sum_i n_i$ for each i, that result in posterior test statistics with lowest variance.

3. Compute two-sample or independence test statistics on this weighted representation of the data to obtain a real-valued scalar $\hat{t}$ that discriminates between the two hypotheses of interest.

4. In practice, a test decision will be made based on a comparison of the computed value $\hat{t}$ with an approximated null distribution obtained by repeated test statistic computation on permuted data representations. If $\hat{t}$ is greater than the α quantile of this approximated null distribution, reject the null hypothesis, otherwise fail to reject.

A.3.3 GP2ST

The test developed by [15] was designed to test the equality of regression functions from observed two-dimensional data $(\mathbf{t}_1, \mathbf{y}_1)$ and $(\mathbf{t}_2, \mathbf{y}_2)$ from two samples. They assume a GP prior on the time series and compute posterior distributions by conditioning on each sample of

observed data. Denote the posterior GPs by f_1 and f_2. With the assumption of gaussianity it follows that $\Delta f := f_1 - f_2$ is also a GP, and evaluations on a fine grid of regular times $\mathbf{t}$ in $[0,1]$ will be multivariate Gaussian with mean denoted $\Delta\mu$ and covariance matrix $\Delta\Sigma$. The hypothesis of equality of data generating processes is then equivalent to testing departures of Δf from the zero function. As a result, the two functions are equal with posterior probability $1-\alpha$ if the credible region for Δf includes the zero vector or, in other words, if:

$$\Delta\mu^T \Delta\Sigma^{-1} \Delta\mu \leq \chi^2_v(1-\alpha). \tag{A.24}$$

$\chi^2_v(1-\alpha)$ is the $(1-\alpha)$-quantile of a χ^2 distribution with v degrees of freedoms and v is the number of positive eigenvalues of $\Delta\Sigma$.

A.3.4 RDC

The Randomized Dependence Coefficient (RDC) measures the dependence between fixed-dimensional random samples X and Y as the largest canonical correlation between k randomly chosen nonlinear projections of their copula transformations. It is formally defined an analyzed in [112], and given by,

$$\hat{\rho}(\mathbf{x},\mathbf{y}) := \sup_{\alpha,\beta} PCC(\alpha^T \Phi_{\mathbf{x}}, \beta^T \Phi_{\mathbf{y}}),$$

where *PCC* is Pearson's correlation coefficient and Φ are nonlinear random projections, such as sine or cosine projections. To apply this function on irregularly observed data, we interpolate as we do with the MMD and HSIC.

We conduct a test using this measure of dependence by repeatedly shuffling the paired time series M times to induce an empirical distribution of $\{\hat{\rho}_m\}_{m=1}^M$ under the null hypothesis of independence. The p-value is then given by $\sum_{m=1}^M \mathbf{1}\{\hat{\rho}_m > \hat{\rho}\}/M$ where $\hat{\rho}$ is the statistic obtained from the observed data.

A.3.5 PCC

The Pearson's correlation coefficient (PCC) is a measure of linear correlation between two variables. It is defined as,

$$\hat{\rho}(\mathbf{x},\mathbf{y}) := \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})}\sqrt{\sum_i (y_i - \bar{y})}}.$$

Similarly to the RDC, we conduct a test using this measure of dependence by repeatedly shuffling the paired time series M times to induce an empirical distribution of $\{\hat{\rho}_m\}_{m=1}^{M}$ under the null hypothesis of independence.

A.3.6 C2ST

We implemented the C2ST with tensorflow in python. We used a RNN with GRU cells in one version and the deepset architecture (with the author's implementation [192]) in the other. The number of samples in each mini-batch is set to 64 the hidden layer size to 10. We optimize model parameters with Adam, learning rate equal to 0.01, and all variables are initialized with Xavier initialization. We use the elu activation functions for each layer and use sigmoid activation for the output layer given that we perform classification.

Both tests proceeds as follows [114]:

Let $\{x_i\}_{i=1}^{n}$ and $\{y_i\}_{i=1}^{n}$ be two samples of observed time series that include their corresponding time points in each case.

1. Construct the data set $\mathcal{D} = \{(x_i, 0)\}_{i=1}^{n} \cup \{(y_i, 1)\}_{i=1}^{n} =: \{(z_i, l_i)\}_{i=1}^{2n}$.
2. Shuffle $\mathcal{D}$ at random and partition into a training set $\mathcal{D}_{tr}$ and a testing set $\mathcal{D}_{te}$.
3. Fit a classifier g on the training set to predict the sample indicator l.
4. Compute test statistic as classification accuracy on $\mathcal{D}_{te}$: $\widehat{t} := \frac{1}{n_{te}} \sum_{(z_i,l_i)\in\mathcal{D}_{te}} \mathbf{1}\{\mathbf{1}\{g(z_i) > 1/2\} = l_i\}$
5. If $\widehat{t}$ is greater that the α quantile of a $\mathcal{N}(1/2, 1/(4n_{te}))$ reject $\mathcal{H}_0$; otherwise accept $\mathcal{H}_0$.

$\mathbf{1}$ is the indicator function.

Appendix B

Appendix to Chapter 3

This appendix provides additional material accompanying Chapter 3: "A Kernel Two-Sample Test for Unbiased Decisions". It is outlined as follows:

- In section B.1 proofs of all propositions and theorems.
- In section B.2 a more detailed description of the example provided in the introduction.
- In section B.3 a detailed description of other tests and our implementations.
- In section B.4 a discussion of computational complexity and possible methods to speed up computations.

B.1 Proofs

In this section we prove the propositions and theorems described in the main body of this chapter.

B.1.1 Proof of Proposition 1

Assume that kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is characteristic and that for all y, $w(x) > 0$ is bounded above by W. A kernel is called characteristic, if the maximum mean discrepancy between probability measures P_{Y^0} and P_{Y^1} induced by k is such that, $\mathrm{MMD}(P_{Y^0}, P_{Y^1}) = 0$ if and only if $P_{Y^0} = P_{Y^1}$. [69] showed that Gaussian kernels are characteristic.

To prove the proposition we exploit the assumption $Y^0, Y^1 \perp\!\!\!\perp T|X$ and recover expectations with respect to the underlying random variables of interest (Y^0, Y^1). Assuming access to the propensity score, $e(x) = p(T = 1|X = x) = \mathbb{E}(I(T = 1)|X = x)$, and for any measurable function of our observed values Y, such as the kernel function k, we have that,

$$\begin{aligned}
\mathbb{E}_{y,y^\star \sim Y|T=1}\left(\frac{k(y,y^\star)}{e(X)e(X^\star)}\right) &= \mathbb{E}_{y,y^\star}\left(\frac{TT^\star k(y,y^\star)}{e(X)e(X^\star)}\right) \\
&= \mathbb{E}_{y,y^\star}\left(\frac{I(T=1)I(T^\star=1)k(y^1,y^{1\star})}{e(X)e(X^\star)}\right) \\
&= \mathbb{E}_{x,x^\star}\left(\mathbb{E}_{y,y^\star}\left(\frac{I(T=1)I(T^\star=1)k(y^1,y^{1\star})}{e(x)e(x^\star)}|y^1,y^{1,\star},x,x^\star\right)\right) \\
&= \mathbb{E}_{y^1,y^{1,\star},x,x^\star}\left(\frac{k(y^1,y^{1\star})}{e(x)e(x^\star)}\mathbb{E}_{T,T^\star}\left(I(T=1)I(T^\star=1)|y^1,y^{1,\star},x,x^\star\right)\right) \\
&= \mathbb{E}_{y^1,y^{1,\star}}\left(k(y^1,y^{1,\star})\right),
\end{aligned}$$

where recall that we use the notation y^1 for a realization of the random variable Y^1. I is the indicator function. This derivation shows that by taking weighted expectations with respect to the observed distribution $Y|T = 1$ we can access expectations with respect to our distribution of interest Y^1. Similar derivations follow for data observed under $Y|T = 0$ using the fact that $\mathbb{E}_{Y|T=0}(\frac{f(Y)}{1-e(X)}) = \mathbb{E}_{Y^0}(f(Y^0))$, for f any measurable function.

Now notice that the $\text{MMD}(Y^0, Y^1)$ between Y^0 and Y^1 is defined in terms of expectations with respect to the random variables Y^0 and Y^1,

$$\text{MMD}(Y^0,Y^1) := \mathbb{E}_{y^0,y^{0,\star}}k(y^0,y^{0,\star}) + \mathbb{E}_{y^1,y^{1,\star}}k(y^1,y^{1,\star}) - 2\mathbb{E}_{y^1,y^0}k(y^0,y^1).$$

Thus with the above derivation we get that each term in the definition of $\text{WMMD}(Y|T = 0, Y|T = 1)$ is equal to each term in the definition of the MMD, which proves the proposition. □

B.1.2 Proof of Theorem 1

Regularity conditions. The following notation is used in the statement on the regularity conditions of Theorem 1. Let $B_n = (b_{imn})$ and $W_n = (W_{ijn})$, for $i, j = 1, ..., n; n, m : 1, 2, ...$ Here W_n is a matrix of weights in $\mathbb{R}^{n\times n}$ and B_n is an orthogonal matrix in $\mathbb{R}^{m\times n}$ such that $B_n^T W_n B_n = \Lambda_n$, where Λ_n is a diagonal matrix with λ_{mn} as the m^{th} diagonal element. Assume

$\lim_{n\to\infty}\lambda_{mn}=\lambda_m$ and let δ_{km} be the dirac delta function with $\delta_{km}=1$ if $k=m$ and zero otherwise. Assume that the following regularity conditions hold,

1. $\max_{1\le i\le n}|b_{imn}|\to 0$ as $n\to\infty$ for each m.
2. $\sum_{i=1}^{n} b_{imn}b_{ikn}\to\delta_{mk}$ as $n\to\infty$ for all m,k.
3. $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ijn}^2\to\sum_{m=1}^{\infty}\lambda_m^2<\infty$.
4. $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ijn}b_{ikn}b_{jkn}\to\lambda_k$ as $n\to\infty$, for all m.

These conditions are sufficient by [46] for a square matrix of data-dependent weights $W=(w_iw_j)$ to be approximately diagonalizable, such that it admits an eigen-decomposition $B^TWB=\Lambda$.

Proof. Recall the definition of the empirical estimate of the WMMD^2,

$$\widehat{\mathrm{WMMD}}^2 := \frac{1}{n(n-1)}\sum_{i\neq j:t_i=t_j=1} w_iw_jk(y_i,y_j)+\frac{1}{m(m-1)}\sum_{i\neq j:t_i=t_j=0}k(y_i,y_j)- \frac{2}{nm}\sum_{i,j:t_i=1,t_j=0} w(x_i)k(y_i,y_j), \tag{B.1}$$

where the (y_i,t_i,x_i) are realization of the random variables (Y,T,X), and have assumed that n observations are made with $T=1$ and m with $T=0$. $w(x_i)=Pr(T_i=1|X_i=x_i)/Pr(T_i=0|X_i=x_i)$ is the density ratio giving the likelihood of an example i being observed under one population with respect to the other. We assume this ratio to be known (for now) and provide approximation bounds for our proposed approximation in Theorem 2 and 3. Our proof is presented in three parts, each one deriving the asymptotic behaviour of each one of the three terms in (B.1).

Note first that we may write the square integrable (centered) kernel k as a weighted sum of product of eigen-functions of the Hilbert-Schmidt operator defined by k [70],

$$k(y_i,y_j)=\sum_{k=1}^{\infty}\alpha_k\psi_k(Y_i)\psi_k(Y_j). \tag{B.2}$$

Consider now the first term in (B.1), it follows that,

$$\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} w(x_i)w(x_j)k(y_i,y_j) = \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} w_{ij}\sum_{k=1}^{\infty}\alpha_k\psi_k(Y_i)\psi_k(Y_j) \tag{B.3}$$

$$= \sum_{k=1}^{\infty}\alpha_k\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} w_{ij}\psi_k(Y_i)\psi_k(Y_j), \tag{B.4}$$

where we have dropped the t_i's in the summation indices and have written $w_{ij} = w(x_i)w(x_j)$ for brevity. Using the degeneracy of k (in the sense that $\mathrm{Var}[\mathbb{E}[k(y,y')]] = 0$), the eigen-functions $\psi_k(Y_i)$, $i = 1,...,n$ are zero mean independent random variables by the independence of the Y_i. Using the above and the regularity conditions, Theorem 1 in [182] yields,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} w_{ij}\psi_k(Y_i)\psi_k(Y_j) \quad \xrightarrow{d} \quad \sum_{m=1}^{\infty}\lambda_m(Z_{km}^2-1), \tag{B.5}$$

where $Z_{km}\sim\mathcal{N}(0,1)$ are $i.i.d.$.

The limiting distribution of the un-weighted term in (B.1) is that of a well-studied U-Statistic whose derivation can be found in Section 5.5.2 of [157].

$$\frac{1}{m}\sum_{i=1:t_i=0}^{m}\sum_{j=1,j\neq i:t_j=0}^{m} k(Y_i,Y_j) \quad \xrightarrow{d} \quad \sum_{k=1}^{\infty}\alpha_k(V_k^2-1). \tag{B.6}$$

The limiting distribution of the cross term in (B.1) follows from a modification of the derivation of Theorem 1 in [46] and is given by,

$$\frac{1}{\sqrt{nm}}\sum_{i=1:t_i=1}^{n}\sum_{j=1:t_j=0}^{n} w'_{ij}\psi_k(Y_i)\psi_k(Y_j) \quad \xrightarrow{d} \quad \sum_{m=1}^{\infty}\lambda'_m Z_{km}V_{km}, \tag{B.7}$$

where the eigenvalues (λ'_m) correspond to those of the eigen-decomposition of the weight matrix W' with $W'_{ij} = w(x_i)$ and where $V_{km}\sim\mathcal{N}(0,1)$ independently of $Z_{km}\sim\mathcal{N}(0,1)$. We prove (*B.7*) below.

We now combine these results. Define $t = m+n$, and assume $\lim_{m,n\to\infty} m/t = \rho_y$ and $\lim_{m,n\to\infty} n/t = \rho_x := (1-\rho_y)$ for fixed $0<\rho_x<1$. Then,

$$t\widehat{\mathrm{WMMD}}^2 \xrightarrow{d} \rho_x^{-1}\sum_{k=1}^{\infty}\alpha_k\sum_{m=1}^{\infty}\lambda_m(Z_{km}^2-1)+\rho_y^{-1}\sum_{k=1}^{\infty}\alpha_k(V_k^2-1)-\frac{2}{\sqrt{\rho_x\rho_y}}\sum_{k=1}^{\infty}\alpha_k\sum_{m=1}^{\infty}\lambda'_m Z_{km}V_{km}. \tag{B.8}$$

In the case that both samples have equal size with total sample size n, we have that under $\mathcal{H}_0$,

$$n\widehat{\mathrm{WMMD}}^2 \xrightarrow{d} \sum_{k=1}^{\infty} \alpha_k \sum_{m=1}^{\infty} \lambda_m (Z_{km}^2 - 1) + \sum_{k=1}^{\infty} \alpha_k (V_k^2 - 1) - 2 \sum_{k=1}^{\infty} \alpha_k \sum_{m=1}^{\infty} \lambda'_m Z_{km} V_{km}. \quad \text{(B.9)}$$

The case of $P \neq Q$, under $\mathcal{H}_1$. The centered kernel k is non-degenerate since its expectation under assumption $\mathcal{H}_1$ is different from 0. The limiting distribution of WMMD can be derived by considering each term in the sum separately. For the first and third terms,

$$(\star) := \frac{1}{n(n-1)} \sum_{i \neq j : t_i = t_j = 1} w(x_i) w(x_j) k(y_i, y_j), \qquad (\star\star) := \frac{2}{mn} \sum_{i,j : t_i = 1, t_j = 0} w(x_i) k(y_i, y_j), \quad \text{(B.10)}$$

we get immediately by Theorem 2.1 from p. 4, [161] that their limiting distributions are normal with mean $\mathbb{E}(\star)$ and variance $\mathrm{Var}(\star)$, and mean $\mathbb{E}(\star\star)$ and variance $\mathrm{Var}(\star\star)$, respectively. The middle term $\frac{1}{m(m-1)} \sum_{i \neq j : t_i = t_j = 0} k(y_i, y_j)$ is an un-weighted U-statistic whose limiting distribution is given by the results in section 5.5 [157]. As above, define $t = m + n$, and assume $\lim_{m,n \to \infty} m/t = \rho_y$ and $\lim_{m,n \to \infty} n/t = \rho_x := (1 - \rho_y)$ for fixed $0 < \rho_x < 1$. Collecting these results, we get under $\mathcal{H}_1$,

$$t^{1/2} \left(\widehat{\mathrm{WMMD}}^2 - \mathrm{WMMD}^2 \right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\mathcal{H}_1}\right), \quad \text{(B.11)}$$

where we write $z = ((y_1, t = 1, x_1), (y_0, t = 0, x_0))$ for the joint sample under the two populations, and $h(z, z^\star) := w(x_1) w(x_1^\star) k(y_1, y_1^\star) + \mathbb{E} k(y_0, y_0^\star) - 2 w(x_1) k(y_1, y_0^\star)$. $\sigma^2_{\mathcal{H}_1} := \mathrm{Var}_z \left(\mathbb{E}_{z^\star} h(z, z^\star) \right)$ [157, 70].

Proof of equation (B.7). The proof is a modification of the result of the convergence of degenerate U statistics on p. 761 in [70] and of the derivation of Theorem 1 in [46].

Consider,

$$T_k := \frac{1}{\sqrt{nm}} \sum_{i=1 : t_i = 1}^{n} \sum_{j=1 : t_j = 0}^{m} w'_{ij} \psi_k(Y_i) \psi_k(Y_j), \quad \text{(B.12)}$$

and define for each k,

$$w^*_{ij} := \sum_{s=1}^{S} \lambda_s b_{isk} b_{jsk}, \qquad T^*_k := \frac{1}{\sqrt{nm}} \sum_{i=1 : t_i = 1}^{n} \sum_{j=1 : t_j = 0}^{n} w'_{ij} \psi_k(Y_i) \psi_k(Y_j). \quad \text{(B.13)}$$

We will start by showing that $\sum_{i=1}^{n}\sum_{j=1}^{m}\left(w_{ij}-w_{ij}^{*}\right)^{2}\to 0$ as $n,m\to\infty$. Note that this implies that $\mathrm{Var}(T_k^{*}-T_k)\to 0$ and thus that the distributions of T_k^{*} and T_k coincide in the limit. We will proceed by showing first the convergence of the sum of squares and then we derive the distribution of T_k^{*}. Using the definitions above, write,

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(w_{ij}-w_{ij}^{*}\right)^{2} &= \sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}^{2}-2\sum_{s=1}^{S}\lambda_s\sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}b_{isk}b_{jsk}+\\ &\sum_{s=1}^{S}\sum_{t=1}^{S}\lambda_s\lambda_t\left(\sum_{i=1}^{n}b_{isk}b_{itk}\right)\left(\sum_{j=1}^{m}b_{jsk}b_{jtk}\right)\\ &=\sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}^{2}-\sum_{s=1}^{S}\lambda_s^{2}-2\sum_{s=1}^{S}\lambda_s\left(\sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}b_{iks}b_{jks}-\lambda_s\right)\\ &+\sum_{s=1}^{S}\sum_{t=1}^{S}\lambda_t\lambda_s\left(\sum_{i=1}^{n}b_{isk}b_{itk}-\delta_{st}\right)\left(\sum_{j=1}^{m}b_{jsk}b_{jtk}-\delta_{st}\right)+\\ &2\sum_{s=1}^{S}\lambda_s^{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}b_{iks}^{2}-1\right),\end{aligned}\tag{B.14}$$

where we have removed the group allocation indices t for clarity. Note here that the first and second term cancel each other by Assumption 1 of the regularity conditions, the third term is $\mathcal{O}(1)$ by Assumption 4 and the fourth and fifth terms are also $\mathcal{O}(1)$ by Assumption 2 and the properties of the dirac delta function.

Consider now T_k^{*} and rewrite it as,

$$T_k^{*}=\sum_{s=1}^{S}\lambda_s\left(\frac{1}{\sqrt{n}}\sum_{i=1:t_i=1}^{n}b_{isk}\psi_k(Y_i)\right)\left(\frac{1}{\sqrt{m}}\sum_{j=1:t_j=0}^{m}b_{jsk}\psi_k(Y_j)\right).\tag{B.15}$$

Define the length K vectors Ψ_n and Ψ'_m having k_{th} entries,

$$\Psi_{kn}=\left(\frac{1}{\sqrt{n}}\sum_{i=1:t_i=1}^{n}b_{isk}\psi_k(Y_i)\right),\qquad \Psi'_{km}=\left(\frac{1}{\sqrt{m}}\sum_{j=1:t_j=0}^{m}b_{jsk}\psi_k(Y_j)\right),\tag{B.16}$$

respectively. These have mean and covariance,

$$\mathbb{E}(\Psi_{kn})=0,\qquad \mathrm{Cov}(\Psi_{kn},\Psi_{k'n})=\begin{cases}\frac{1}{m}\sum_{i=1}^{n}b_{isk}^{2}=1, & \text{if}\quad k=k'\\ 0, & \text{otherwise.}\end{cases}\tag{B.17}$$

Moreover, the vectors Ψ_n and Ψ'_m are independent. The results (B.7) then holds by the Lindberg-Levy Central Limit Theorem [157], Theorem 1.9.1A. □

B.1.3 Proof of Theorem 2

We assume that for increasing sample size, as $n,m \to \infty$, we can approximate arbitrarily well the density ratio $w(x)$, for all x in our training data. This is justified by the following Lemma,

Lemma 1 (Lemma 1.4 [74]) *Let $w(x_i) \in [0,B]$ be the optimal weight in the population sense, $Pr(T_i = 1|x_i) = w(x_i)Pr(T_i = 0|x_i)$. Assume we draw n samples from $X|T = 1$ and m samples from $X|T = 1$ independently and that $||\phi(x)|| \leq R$. Then, with probability at least $1-\delta$,*

$$\left|\left|\frac{1}{n}\sum_{i=1:t_i=1}^{n} w(x_i)\phi(x_i) - \frac{1}{m}\sum_{j=1:t_j=0}^{m}\phi(x_j)\right|\right| \leq \left(1+\sqrt{2\log\frac{2}{\delta}}\right)R\sqrt{\frac{B^2}{n}+\frac{1}{m}}. \tag{B.18}$$

Note that because the optimization problem is convex the choice of $\hat{w}(x) := Pr(T_i = 1|x)/Pr(T_i = 0|x)$ uniquely minimizes the objective function with value 0, see Lemma 1.3, [74]. Thus by the argument above, we may assume that for increasing sample size, as $n,m \to \infty$, $\hat{w}(x) \to w(x)$, for all x in the common support of the distributions $Pr(T_i = 1|x)$ and $Pr(T_i = 0|x)$.

Consider the first terms of $\widehat{\text{WMMD}}^2(;\hat{w})$ and $\widehat{\text{WMMD}}^2(;w)$, that denote the empirical WMMD^2 with estimated and true weights w respectively,

$$\hat{K}_{n,m} := \sum_{i=1:t_i=1}^{n}\sum_{j=1,j\neq i:t_j=1}^{m}\hat{w}_{ij}k(y_i,y_j), \quad \text{and} \quad K_{n,m} := \sum_{i=1:t_i=1}^{n}\sum_{j=1,j\neq i,t_j=1}^{m} w_{ij}k(y_i,y_j). \tag{B.19}$$

It holds that $\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{m}\left(\hat{w}_{ij}-w_{ij}\right)^2 \to 0$ as $n,m \to \infty$ by the arguments above. This implies that $\text{Var}(\hat{K}_{n,m} - K_{n,m}) \to 0$ and $E(|\hat{K}_{n,m} - K_{n,m}|^2) \to 0$ which means that $\hat{K}_{n,m} - K_{n,m}$ converges to 0 in L_2, and hence in distribution. The distributions of $\hat{K}_{n,m}$ and $K_{n,m}$ coincide in the limit.

The same derivations apply for the other two terms in the definition of $\widehat{\text{WMMD}}^2$. Therefore we conclude that $\widehat{\text{WMMD}}^2$ with estimated weights has the same asymptotic null and alternative distribution as $\widehat{\text{WMMD}}^2$ with known weights. In particular, asymptotically, its false positive rate is α and its power converges to 1.

B.1.4 Proof of Theorem 3

We prove Theorem 3 by first stating and proving several Lemmas which bound the different terms of the inequality of interest.

Lemma 2 *In addition to the conditions of Lemma 1, assume there exists some $\hat{w}_i$, the empirical counterparts of the population weights estimated by matching kernel mean embeddings, such that,*

$$\left\| \frac{1}{n} \sum_{i=1:t_i=1}^{n} \hat{w}_i \phi(x_i) - \frac{1}{m} \sum_{j=1:t_j=0}^{m} \phi(x_j) \right\| \leq \varepsilon. \tag{B.20}$$

Then,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} w_i \phi(x_i) - \frac{1}{n} \sum_{i=1}^{n} w(x_i) \phi(x_i) \right\| \leq \varepsilon + \left(1 + \sqrt{2 \log \frac{2}{\delta}} \right) R \sqrt{\frac{B^2}{n} + \frac{1}{m}}. \tag{B.21}$$

Proof. Note that by using Lemma 1 and the triangle inequality we immediately get,

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1:t_i=1}^{n} \hat{w}_i \phi(x_i) - \frac{1}{n} \sum_{i=1:t_i=1}^{n} w_i \phi(x_i) \right\| &\leq \left\| \frac{1}{n} \sum_{i=1:t_i=1}^{n} \hat{w}_i \phi(x_i) - \frac{1}{m} \sum_{j=1:t_j=0}^{m} \phi(x_j) \right\| \\ &\quad + \left\| \frac{1}{n} \sum_{i=1:t_i=1}^{n} w_i \phi(x_i) - \frac{1}{m} \sum_{j=1:t_j=0}^{m} \phi(x_j) \right\| \\ &\leq \varepsilon + \left(1 + \sqrt{2 \log \frac{2}{\delta}} \right) R \sqrt{\frac{B^2}{n} + \frac{1}{m}}. \end{aligned} \tag{B.22}$$

□

Lemma 3 *Let $\widehat{WMMD}(w)$ be the weighted estimator of the MMD given i.i.d. distorted samples as defined in equation (B.1) with known (population) weights w, and similarly define $\widehat{WMMD}(\hat{w})$ with weights $\hat{w}$ estimated by matching the empirical kernel mean embeddings of the distorted samples. Then, given the conditions of Lemmas 1 and 2,*

$$\left| \widehat{\mathrm{WMMD}}^2(\hat{w}) - \widehat{\mathrm{WMMD}}^2(w) \right| \leq 2R(B+1) \left(\varepsilon + \left(1 + \sqrt{2 \log \frac{2}{\delta}} \right) R \sqrt{\frac{B^2}{n} + \frac{1}{m}} \right). \tag{B.23}$$

Proof. Consider expanding the estimators,

$$\left|\widehat{\mathrm{WMMD}}^2(\hat{w}) - \widehat{\mathrm{WMMD}}^2(w)\right| = \frac{1}{n(n-1)}\sum_{i\neq j}\hat{w}_i\hat{w}_j k(y_i,y_j) - \frac{1}{n(n-1)}\sum_{i\neq j} w_i w_j k(y_i,y_j) - \left(\frac{2}{nm}\sum_{i,j}\hat{w}_i k(y_i,y_j) - \frac{2}{nm}\sum_{i,j} w(x_i)k(y_i,y_j)\right). \tag{B.24}$$

Note that the U-statistic in y cancel since these do not involve the weights.

First and second terms. We can bound the first and second terms as follows,

$$\frac{1}{n(n-1)}\sum_{i\neq j}\hat{w}_i\hat{w}_j k(y_i,y_j) - \frac{1}{n(n-1)}\sum_{i\neq j} w_i w_j k(y_i,y_j) \tag{B.25}$$

$$= \frac{1}{n(n-1)}\sum_{i\neq j}\hat{w}_i\hat{w}_j \left\langle \psi(y_i), \psi(y_j)\right\rangle - \frac{1}{n(n-1)}\sum_{i\neq j} w_i w_j \left\langle \psi(y_i), \psi(y_j)\right\rangle \tag{B.26}$$

$$= \Bigg| \left\langle \frac{1}{n}\sum_{i=1}^n \hat{w}_i\psi(y_i) - \frac{1}{n}\sum_{i=1}^n w(x_i)\psi(y_i), \frac{1}{n-1}\sum_{j=1,j\neq i}^m \hat{w}_j\psi(y_j)\right\rangle + \left\langle \frac{1}{n}\sum_{i=1}^n \hat{w}_i\psi(y_i) - \frac{1}{n}\sum_{i=1}^n w(x_i)\psi(y_i), \frac{1}{n-1}\sum_{j=1,j\neq i}^m w(x_j)\psi(y_j)\right\rangle \Bigg| \tag{B.27}$$

$$\leq \left| \left\langle \frac{1}{n}\sum_{i=1}^n \hat{w}_i\psi(y_i) - \frac{1}{n}\sum_{i=1}^n w(x_i)\psi(y_i), \frac{1}{n-1}\sum_{j=1,j\neq i}^m \hat{w}_j\psi(y_j)\right\rangle \right| + \left| \left\langle \frac{1}{n}\sum_{i} \hat{w}_i\psi(y_i) - \frac{1}{n}\sum_{i} w(x_i)\psi(y_i), \frac{1}{n-1}\sum_{j=1,j\neq i}^m w(x_j)\psi(y_j)\right\rangle \right| \tag{B.28}$$

$$\leq 2BR\left(\varepsilon + \left(1+\sqrt{2\log\frac{2}{\delta}}\right)R\sqrt{\frac{B^2}{n}+\frac{1}{m}}\right), \tag{B.29}$$

where $\psi(y) := k(y,\cdot)$. Note that we have omitted the group allocation indices, these should be clear however from the i and j indices. The second equality follows by adding and subtracting $\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j=1,j\neq i}^m w(x_i)\hat{w}_j\langle\psi(y_i),\psi(y_j)\rangle$ which factorizes into the given expression. The second to last inequality follows from the triangle inequality and the last inequality follows from the properties of norms and the results derived in Lemmas 1 and 2.

Third and fourth terms. The third and fourth terms (in brackets) are derived similarly and satisfy the following bounds,

$$\frac{2}{nm}\sum_{i,j}\hat{w}_i k(y_i,y_j) - \frac{2}{nm}\sum_{i,j} w(x_i)k(y_i,y_j) \tag{B.30}$$

$$= \frac{2}{nm}\sum_{i,j}\hat{w}_i \left\langle \psi(y_i),\psi(y_j)\right\rangle - \frac{2}{nm}\sum_{i,j} w(x_i)\left\langle \psi(y_i),\psi(y_j)\right\rangle \tag{B.31}$$

$$= \left| \frac{1}{n}\sum_{i=1}^{n}\hat{w}_i \left\langle \psi(y_i), \frac{2}{m}\sum_{j=1}^{m}\psi(y_j)\right\rangle - \frac{1}{n}\sum_{i=1}^{n} w(x_i)\left\langle \psi(y_i), \frac{2}{m}\sum_{j=1}^{m}\psi(y_j)\right\rangle \right| \tag{B.32}$$

$$= \left| \left\langle \frac{1}{n}\sum_{i=1}^{n}\hat{w}_i\psi(y_i) - \frac{1}{n}\sum_{i=1}^{n} w(x_i)\psi(y_i), \frac{2}{m}\sum_{j=1}^{m}\psi(y_j)\right\rangle \right| \tag{B.33}$$

$$\leq 2R\left(\varepsilon + \left(1+\sqrt{2\log\frac{2}{\delta}}\right) R\sqrt{\frac{B^2}{n}+\frac{1}{m}}\right), \tag{B.34}$$

where the last inequality follows from the properties of norms and the results derived in Lemmas 1 and 2.

Finally, collecting the two bounds the lemma follows. □

Lemma 4 *Let $\widehat{WMMD}(w)$ be the weighted estimator of the MMD given i.i.d. distorted samples as defined in equation (B.1) with known (population) weights w, and maximum kernel value R. Assume that $1 \leq w \leq B$ for all $x \in \mathcal{X}$. Then, with probability at least $1-\delta$,*

$$\left|\widehat{\mathrm{WMMD}}^2(w) - \mathrm{MMD}^2\right| \leq R(B+1)^2\sqrt{\frac{1}{2m_2}\log\frac{1}{\delta}}, \tag{B.35}$$

where $m_2 := \lfloor m/2 \rfloor$.

Proof. Assuming the kernel $k(\cdot,\cdot)$ is bounded between 0 and R and the weights w bounded between 0 and B, we can infer function bounds such that $-2BR \leq wk(y_i,x_j) \leq R(B^2+1)$. By Theorem 10 in [70] which results from an application of the large deviation bound on U statistics due to Hoeffding we have that,

$$p\left(\left|\widehat{\mathrm{WMMD}}^2(w) - \mathrm{MMD}^2\right| > e\right) \leq \exp\left\{\frac{-2e^2 m_2}{R^2(B+1)^4}\right\}. \tag{B.36}$$

Define $\delta = \exp\left\{\frac{-2e^2 m_2}{R^2(B+1)^4}\right\}$. Thus, with probability $1-\delta$,

$$\left|\widehat{\mathrm{WMMD}}^2(w) - \mathrm{MMD}^2\right| \leq R(B+1)^2\sqrt{\frac{1}{2m_2}\log\frac{1}{\delta}}, \tag{B.37}$$

where $m_2 := \lfloor m/2 \rfloor$. □

We are ready to prove Theorem 3. This will be a straightforward combination of the lemmas given above.

Proof of Theorem 3. Let $\widehat{\mathrm{WMMD}}(\hat{w})$ be the weighted estimator of the MMD given *i.i.d.* distorted samples as defined in (B.1) with estimated weights $\hat{w}$. Assume conditions on Lemmas 1,2,3 and 4 above hold and that there exists an $\varepsilon > 0$ such that,

$$\left|\left|\frac{1}{n}\sum_{i=1}^{n}\hat{w}_i\phi(x_i) - \frac{1}{m}\sum_{i=1}^{m}\phi(x_i)\right|\right| \leq \varepsilon. \tag{B.38}$$

We may decompose the absolute difference between our weighted approximation using distorted samples and the population MMD as follows,

$$\begin{aligned}&\left|\widehat{\mathrm{WMMD}}^2(\hat{w}) - \mathrm{MMD}^2\right| \\ &\qquad \leq \left|\widehat{\mathrm{WMMD}}^2(\hat{w}) - \widehat{\mathrm{WMMD}}^2(w)\right| + \left|\widehat{\mathrm{WMMD}}^2(w) - \mathrm{MMD}^2\right|.\end{aligned} \tag{B.39}$$

Then using Lemma 3 to bound the first term and Lemma 4 to bound the second term, we get that with probability at least $1-\delta$,

$$\begin{aligned}&|\widehat{\mathrm{WMMD}}^2(\hat{w}) - \mathrm{MMD}^2| \leq \\ &\qquad R(B+1)\left(2\varepsilon + 2\left(1+\sqrt{2\log\frac{2}{\delta}}\right)R\sqrt{\frac{B^2}{n}+\frac{1}{m}} + (B+1)\sqrt{\frac{1}{2m_2}\log\frac{1}{\delta}}\right),\end{aligned} \tag{B.40}$$

where $m_2 := \lfloor m/2 \rfloor$. □

B.2 Details on the introductory example

The example is used to illustrate the need for adjusting for confounding variables. For a total of 500 individuals we generated random education data X by sampling from a uniform distribution between 0 and 10, from which we derived the post-intervention income Y^0 and Y^1 by simply adding a standard random Gaussian noise variable to these values (in this case $\mathcal{H}_0$ holds: the distributions are equal). We generated male $T = 1$ and female $T = 0$ data, our two populations ($S = 1$), by selectively removing with probability 0.5 females with education level higher than 5 ($Pr(T = 0|X > 5) \approx 0.33$), and removing with probability 0.5 males with education level lower than 5 ($Pr(T = 0|X < 5) \approx 0.66$). We end up with approximately 150 individuals in each group, males with higher education levels than females on average. Observe that the underlying generating process is the same in both populations, only the marginal distribution of the education level changes. As is natural, a two-sample test that overlooks the differences in education will reject the hypothesis of equal data generating process for the income.

B.3 Description and implementation of tests

B.3.1 Hyperparameter selection for high power

The population quantity $\mathrm{WMMD} = 0$ if and only if the distributions under consideration are equal, for any choice of kernel hyperparameters. With finite sample size n, decisions must rely on inference based on the empirical WMMD, and some hyperparameters will give higher power than others. A popular strategy is to set the bandwidth σ of the Gaussian kernel to the median squared pairwise distance between input data, but can be sub-optimal when the scale of the difference between populations differs from the scale of the difference within populations themselves. Instead, we follow the approaches of [169, 86] and choose σ so as to maximize the test power, i.e. the probability of rejecting $\mathcal{H}_1$ when it is false.

Proposition (Approximate power of test statistic). *Under $\mathcal{H}_1$, for large n and fixed r, the test power $Pr(n\widehat{\mathrm{WMMD}}^2 > r) \approx 1 - \Phi(\frac{r}{\sqrt{n}\sigma_{\mathcal{H}_1}} - \sqrt{n}\frac{\mathrm{WMMD}^2}{\sigma_{\mathcal{H}_1}})$, where Φ denotes the cumulative distribution function of the standard normal distribution, and $\sigma_{\mathcal{H}_1}$ is defined as in Theorem 1.*

Assume that n is sufficiently large. Following the same argument as in [86], in $\frac{r}{\sqrt{n}\sigma_{\mathcal{H}_1}} - \frac{\mathrm{WMMD}^2}{\sigma_{\mathcal{H}_1}}$, we observe that the first term $\frac{r}{\sqrt{n}\sigma_{\mathcal{H}_1}} = \mathcal{O}(n^{-1/2})$ goes to 0 as $n \to \infty$ because $\sigma^2_{\mathcal{H}_1} = \mathcal{O}(n^{-1})$, while the second term, $\sqrt{n}\frac{\mathrm{WMMD}^2}{\sigma_{\mathcal{H}_1}} = \mathcal{O}(n^{1/2})$, dominates the first one for large n. Thus, the

parameters that maximize the test power are given by $\theta^* = \text{argmax}_\theta \; p(n\widehat{\text{WMMD}}^2 > r) \approx \frac{\text{WMMD}^2}{\sigma_{\mathcal{H}_1}}$. Since WMMD and $\sigma_{\mathcal{H}_1}$ are unknown, to maintain the validity of the hypothesis test we divide the sample into a training set, used to compute $\frac{\widehat{\text{WMMD}}^2}{\hat{\sigma}_{\mathcal{H}_1}}$ and choose the kernel, and a testing set used to perform the final hypothesis test with the learned kernel. The empirical estimate of the variance $\hat{\sigma}_{\mathcal{H}_1}$ that appears in our objective is approximated up to second order terms, similarly to [169].

B.3.2 B-Test: a modification that uses propensity scores

An alternative to the weighted MMD test is a B-test (block-based test): the idea is to break the data into homogeneous blocks by stratifying subjects into mutually exclusive subsets based on their estimated propensity score. Recall that the propensity score is defined as $e(x) := Pr(T = 1|X)$, the probability of group assignment given confounding variables. After this stage, we compute a two sample test statistic on each block, and average these quantities to obtain the test statistic.

More specifically, subjects are ranked according to their estimated propensity score and then stratified into subsets based on previously defined thresholds of the estimated propensity score. Because population assignment is essentially at random for individuals with the same propensity value, we expect mean comparisons within this group to be unbiased. [142] showed that stratification based on the propensity score will balance x, in the sense that within strata homogeneous in $e(x) = Pr(T = 1|x)$, the distribution of x will be equal in the two populations.

For an individual block, laying on the main diagonal and starting at position $(i-1)B+1$, the statistic $\eta(i)$ is calculated as,

$$\eta(i) := \frac{1}{\binom{B}{2}} \sum_{a=(i-1)B+1}^{iB} \sum_{b=(i-1)B+1 \neq a}^{iB} h(y_{a,0}, y^\star_{b,0}, y_{a,1}, y^\star_{b,1}), \tag{B.41}$$

where $h(y_0, y_0^\star, y_1, y_1^\star) = k(y_0, y_0^\star) + k(y_1, y_1^\star) - k(y_0, y_1^\star) - k(y_0^\star, y_1)$, y_0 is a sample from $Y|T = 0$, y_1 a sample from $Y|T = 1$ and superscript $\star$ denotes an independent copy. The overall test statistic is then,

$$\eta = \frac{B}{n} \sum_{i=1}^{\frac{n}{B}} \eta(i). \tag{B.42}$$

The choice of B determines the accuracy of the balancing procedure and computation time - at one extreme is exact matching based on the propensity score and the linear-time MMD suggested by [70] where we have $n/2$ blocks of size $B = 2$, and at the other extreme is the unbalanced and usual full MMD with 1 block of size n. We chose as a default to divide both populations into $\sqrt{n}$ blocks as proposed in [194].

B-test of [194] assumes that $B \to \infty$ together with n, which implies that the statistic $\hat{\eta}$ defined in (B.42) under the null distribution satisfies,

$$\sqrt{nB}\hat{\eta} \to_d \mathcal{N}(0, 4\sigma^2), \tag{B.43}$$

where $\sigma^2 = E_{X,X'}(k(X,X')^2) + (E_{X,X'}k(X,X'))^2 - 2E_X[(E_{X'}k(X,X')^2]$ that can be estimated directly or by considering the empirical variance of the statistics computed within each of the blocks.

B.3.3 ANCOVA

Analysis of covariance (ANCOVA) are a general statistical procedure derived from a general linear model which blend ANOVA and regression. Conventionally, ANCOVA evaluates whether the means of a dependent variable are equal across levels of a categorical independent variable often called a treatment, while statistically controlling for the effects of other continuous variables that are not of primary interest, that is confounders. In existing implementations [172] these suffer from a number of limitations such as the assumption of an underlying linear feature/outcome mapping and normality of residuals.

In our implementation we proceed as follows. We fit a Random Forest regression model on the confounding variables to approximate the outcome variable Y. Since in our experiments we consider Y to be multivariate, we fit a different regression model for each dimension of Y. We interpret the resulting residuals as being independent of confounders given group assignments and use those to proceed with testing. Because of the computational burden of this procedure, we fit the well-known Hotelling T^2 test [78] on the residuals to decide whether Y^0 and Y^1 share the same generating process up to confounding variables.

B.4 Computational complexity

The computational complexity of the WMMD^2 is quadratic in the number of samples due to the need to compute the Kernel matrix, similarly to the plain implementation of the MMD^2. When permutations are chosen to approximate the null distribution, this procedure can be overly time consuming for large data sets. Below we briefly describe existing approximations that can be used with the WMMD^2 to speed up computations.

- Gamma approximation to the null [72]. This procedure consist of using a two-parameter Gamma distribution that we fit by matching the first and second moments of the empirical MMD^2. Such approximations can be accurate in practice and much faster, although they remain heuristics with no consistency guarantees.
- Linear time test [70]. Another alternative would be to randomly subsample the data such as to make the computational complexity linear in the original number of samples. The drawback is that power is often overly reduced as a result.
- Kernel matrix approximation with low-dimensional random features [136]. To accelerate the computation of the kernel matrix, one may map the input data to a randomized low-dimensional feature space and compute inner products based on these representations. [136] showed that by projecting unto a suitable basis the inner products of the transformed data are approximately equal to those in the feature space of a user specified shift-invariant kernel.

Appendix C

Appendix to Chapter 4

This appendix provides additional material accompanying Chapter 4: "Conditional Independence Testing using Generative Adversarial Networks". It is outlined as follows:

- In sections C.1 and C.2 further experiments relating to. hyperparameters and computational complexity
- In section C.3 the proofs of all theoretical statements.
- In section C.4 implementation details for all methods.
- In section C.5 further details on the genetic data.

C.1 Discussion on hyperparameter choice

As ground truth information on variable relationships is rarely available, choosing hyperparameters is challenging. In this section we analyze the GCIT's performance as a function of hyperparameter configurations of the GCIT and discuss an approximate procedure to guide hyperparameter optimization on a validation set.

C.1.1 Choice of statistic ρ

We start by analyzing potential choices for the summary statistic $\rho : (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}) \times (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}) \to \mathbb{R}$ that summarizes the generated and observed samples into a real-valued scalar.

Different choices for ρ encode in more or less detail the distributional differences in samples and thus we can expect them to influence the resulting performance of the test. We considered the following distance and correlation measures between two samples:

- The Maximum mean discrepancy (MMD) is defined as the largest difference between the mean function values on two samples in a reproducing kernel Hilbert space. When MMD is large, the samples are likely from different distributions. For a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, a consistent empirical estimate of the MMD is given by [70],

$$\hat{\rho}(\mathbf{x},\mathbf{y}) := \frac{1}{n^2}\sum_{i,j} k(x_i,x_j) + \frac{1}{n^2}\sum_{i,j} k(y_i,y_j) - \frac{2}{n^2}\sum_{i,j} k(x_i,y_j).$$

- The Pearson's correlation coefficient (PCC) is a measure of linear correlation between two variables. It is defined as,

$$\hat{\rho}(\mathbf{x},\mathbf{y}) := \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})}\sqrt{\sum_i (y_i-\bar{y})}}.$$

- The distance correlation (DC) measures both linear and nonlinear association between two random variables or random vectors. It is defined as,

$$\hat{\rho}(\mathbf{x},\mathbf{y}) := \frac{dCov(\mathbf{x},\mathbf{y})}{\sqrt{dVar(\mathbf{x})dVar(\mathbf{y})}}.$$

- The Kolmogorov-Smirnov statistic (KS) is defined as the sup-norm between cumulative distribution functions of two samples as follows,

$$\hat{\rho}(\mathbf{x},\mathbf{y}) := \sup_w |F_x^{(n)}(w) - F_y^{(n)}(w)|,$$

  where $F_x^{(n)}$ and $F_y^{(n)}$ are the empirical distribution functions of X and Y samples respectively.

- The Randomized Dependence Coefficient (RDC) measures the dependence between random samples X and Y as the largest canonical correlation between k randomly chosen nonlinear projections of their copula transformations. It is formally defined an analyzed in [112].

$$\hat{\rho}(\mathbf{x},\mathbf{y}) := \sup_{\alpha,\beta} PCC(\alpha^T \Phi_{\mathbf{x}}, \beta^T \Phi_{\mathbf{y}}),$$

where *PCC* is Pearson's correlation coefficient and Φ are nonlinear random projections, such as sine or cosine projections. See [112] for more details.

We tested the above metrics with simulated data under setting **(3)** described in the main chapter. Type I error and power results for the GCIT implemented with each one of the above choices for ρ are given in Figure C.1. Finer differences are given by the MMD, the RDC or the DC that all consider non-linear relationships between variables; we see in the power computations in the right column that this results in higher power of the GCIT since the underlying data generating mechanism in non-linear. However, these statistics will also encode spurious differences between samples when the null $\mathcal{H}_0$ is in fact true, resulting in higher type I error. We can see this behaviour in the type I error results on the panels in the left column. The PCC, for example, that encodes only linear differences between samples is more robust to type I error.

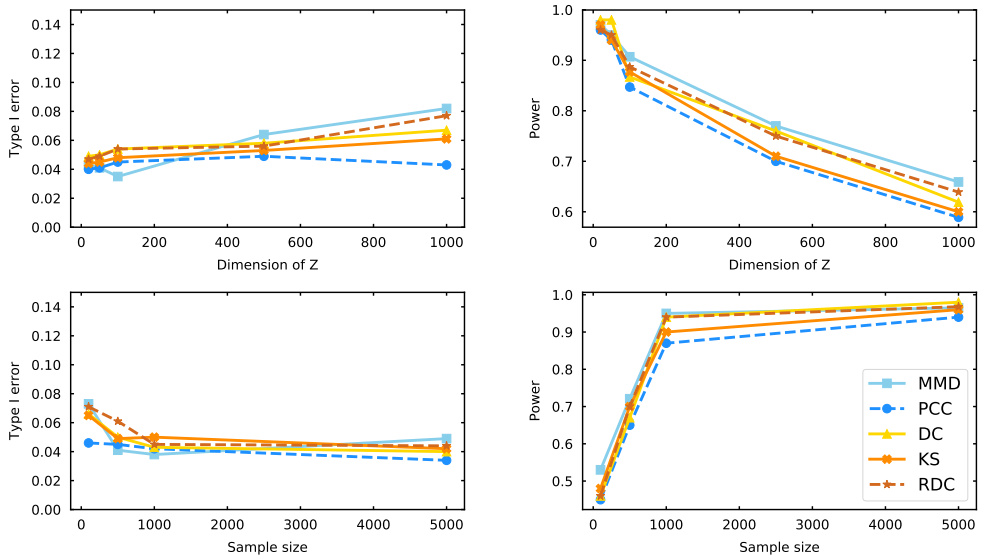


Figure C.1: Power and type I error results for different choices of ρ.

Remark on the robustness of the GCIT for practical applications. *Our test does depend to some extent on the hyperparameter configurations of both λ and ρ. Recall that no ground truth is available to optimize hyperparameters using conventional methods, but we argue that the following procedure can be used to guide hyperparameter selection. We consider artificially inducing conditional independence $(X \perp\!\!\!\perp Y|Z)$ by permuting variables X and Y such as to preserve the marginal dependence in (X,Z) and (Y,Z), as in [50] (further details are also described in our related work section). On this data, a well calibrated test is expected to produce uniformly distributed p-values, i.e. the empirical distribution of p-values should be approximately uniform. Our recommendation would be to choose GCIT's hyperparameters*

with lowest Kolmogorov-Smirnov statistic in comparison to the uniform distribution. This ensures the resulting test produces "well-behaved" p-values and thus prevents to some extent p-value cheating. We will discuss this further in the revised manuscript, thank you for raising this point.

C.2 Further experiments and complexity analysis

In this section we present results on the type I error of the GCIT and all baseline algorithms for the synthetic simulations considered in the main body of this chapter, and analyze computational complexity as a function of sample size and data dimensionality.

C.2.1 Type I error versus dimensionality of Z

Next we show in Figure C.2 type I error as a function of dimensionality of Z for each one of the three synthetic simulations considered in the main body of this chapter. We observe that in all cases, type I error is approximately controlled at the chosen level $\alpha = 0.05$ when the distributional assumptions underlying each method holds. This is not the case otherwise, the CRT fails to control type I error in the non-linear setting when a Gaussian approximation to the joint distribution of the variables is not appropriate.

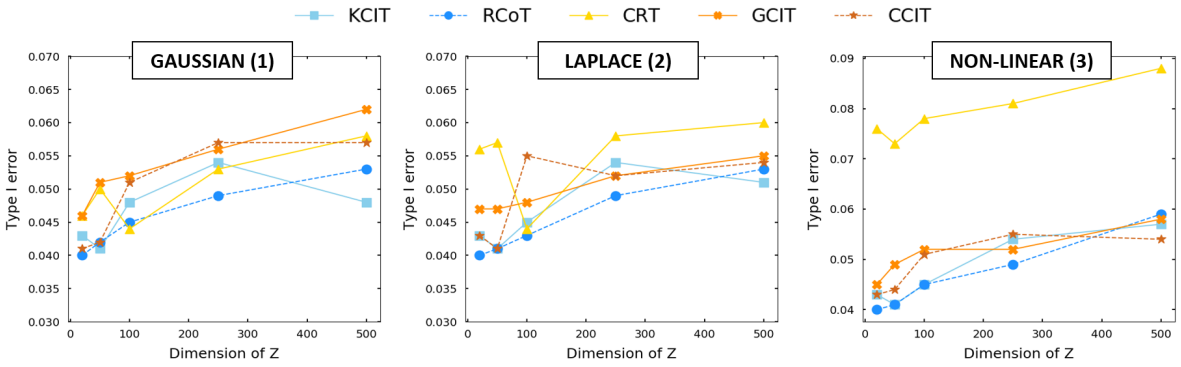


Figure C.2: Type I error results for the synthetic simulations.

C.2.2 Computational complexity analysis

We give in Figure C.3 the run times in seconds of all algorithms for a single conditional independence test for data generated under setting **(1)** in the main body of this chapter. We

vary both the number of samples (fixing the dimension of Z to 100) and the dimensionality of Z (fixing the sample size to 100). The GCIT scales very well with both sample size and conditioning set size, even if each iteration requires training a new GAN. In contrast, the running times of KCIT for sample sizes above 1000 and those of CCIT in higher dimensional samples are prohibitive in practice.

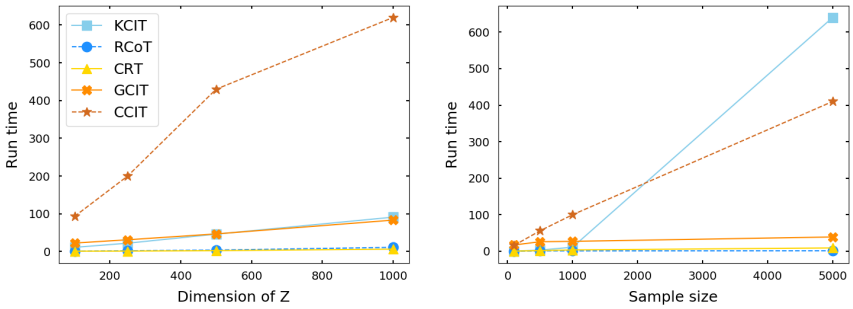


Figure C.3: Running times in seconds as a function of sample size and dimension of Z.

C.2.3 Sensitivities to sample size and stability of generated p-values

We investigate the influence of sample size on the three leftmost panels of Figure C.4. The GCIT, as well as most competing tests, have slightly higher type I error in low sample sizes but control type I error successfully with 500 samples or more. In terms of power, our experiments show that we can expect the GCIT to outperform competing tests with 500 samples or more (for dimension of $Z = 100$). Next, we investigate the stability of p-values as a function of sample size; the variance of the empirical p-values quickly drops to 0. This means that for say 500 samples, we can expect the p-values of two independently trained GCITs to be within 0.005 of each other with approximately 95% confidence. The last panel on the right illustrates how quickly the p-value approximation (eq. 3 in the main body of this chapter) converges to its population quantity as a function of the number of samples used to compute the approximation i.e. M in eq. 3. The convergence should be at least of order $M^{-1/2}$ by the central limit theorem.

C.3 Theoretical results

Proof of Proposition 1. A sequence of random variables is said to be exchangeable if its distribution is invariant under variable permutations. We make use of the "representation theorem" for exchangeable sequences of random variables, first stated by de Finetti and extended

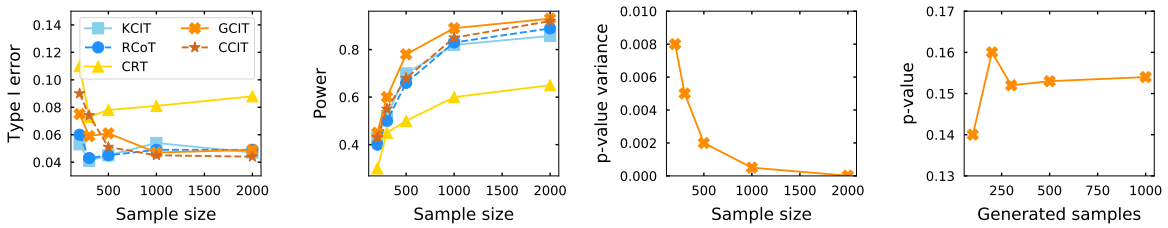


Figure C.4: **Leftmost and middle-left panel**: Type I error and power as a function of sample size for data generated under scenario (3) with dimensionality of Z set to 100; **Middle-right panel**: Empirical p-value variance of the GCIT as a function of sample size (computed by generating 100 p-values for each GAN trained on data with the specified size); **Rightmost panel**: Illustration of the convergence of the GCIT's p-values as a function of generated samples.

by Diaconis and Freedman for finite sequences [45, 47]. They show that every sequence of conditionally $i.i.d.$ random variables can be considered as a sequence of exchangeable random variables. With our definition of the generator we start from $i.i.d.$ sequence of noise random variables $\{V_m\}_{m=1}^M$ and define, for every m, $\tilde{X}^{(m)} = \phi(Z, V_m)$ where Z is a random variable independent of V_m and ϕ is a measurable function, such as a neural network in our case. By construction, the resulting random sequence of data sets $(\tilde{X}^{(m)}, Y, Z)_{m=1}^M$ is exchangeable and therefore also the sequence of statistics $(\rho_i)_{i=1}^M$ (measurable functions of $(\tilde{X}^{(m)}, Y, Z)_{m=1}^M$) is exchangeable. □

The theoretical results that follow are proven only for the version of the generator loss given in equation (5) in the main body of this chapter, $\mathcal{L}_G(D) := \mathbb{E}_{\tilde{x}\sim\hat{q}_{\mathcal{H}_0}} D_\eta(\tilde{x}) - \mathbb{E}_{x\sim q_{\mathcal{H}_0}} D_\eta(x)$ though we do believe that the theorem holds more generally with the addition of the power maximizing procedure - this is backed up by our empirical results demonstrating Type I error control while using the power maximizing procedure. We prove the bound on the excess Type I error in two parts. First we show in the following lemma that an optimal discriminator exists, and second we prove the bound on the Type I error.

Lemma 1 *An optimal discriminator D^* minimizing $\mathcal{L}_D := \mathbb{E}_{x\sim q_{\mathcal{H}_0}} D(x) + \mathbb{E}_{x\sim \hat{q}_{\mathcal{H}_0}}(1-D(x))$ over all measurable functions D such that $D \in (0,1)$ exists and it is given by,*

$$D^* = \frac{1}{2}\text{sign}\left(q_{\mathcal{H}_0} - \hat{q}_{\mathcal{H}_0}\right) + \frac{1}{2}.$$

Proof. To see this note first that the $sign(x)$ function is defined as $+1$ or -1 depending on the sign of x. Then,

$$\begin{aligned}
\mathcal{L}_{D^*} &= 1 + \mathbb{E}_{x\sim q_{\mathcal{H}_0}} D^*(x) - \mathbb{E}_{x\sim \hat{q}_{\mathcal{H}_0}} D^*(x) && \text{(C.1)}\\
&= 1 + \int_{\mathcal{X}} q_{\mathcal{H}_0}(x)\frac{1}{2}\text{sign}\left(q_{\mathcal{H}_0} - \hat{q}_{\mathcal{H}_0}\right)dx - \int_{\mathcal{X}} \hat{q}_{\mathcal{H}_0}(x)\frac{1}{2}\text{sign}\left(q_{\mathcal{H}_0} - \hat{q}_{\mathcal{H}_0}\right)dx && \text{(C.2)}\\
&= 1 - \frac{1}{2}\int_{x:\hat{q}_{\mathcal{H}_0}(x) - q_{\mathcal{H}_0}(x) > 0} q_{\mathcal{H}_0}(x) - \hat{q}_{\mathcal{H}_0}(x)dx - \frac{1}{2}\int_{x:\hat{q}_{\mathcal{H}_0}(x) - q_{\mathcal{H}_0}(x) < 0} \hat{q}_{\mathcal{H}_0}(x) - q_{\mathcal{H}_0}(x)dx && \text{(C.3)}\\
&= 1 - \frac{1}{2}\int_{\mathcal{X}} |\hat{q}_{\mathcal{H}_0}(x) - q_{\mathcal{H}_0}(x)|dx && \text{(C.4)}\\
&= 1 - \frac{1}{2}\sup_{||D||_\infty \leq 1} \mathbb{E}_{x\sim \hat{q}_{\mathcal{H}_0}} D(x) - \mathbb{E}_{x\sim q_{\mathcal{H}_0}} D(x) && \text{(C.5)}\\
&= 1 - \sup_{||D||_\infty \leq 1/2} \mathbb{E}_{x\sim \hat{q}_{\mathcal{H}_0}} D(x) - \mathbb{E}_{x\sim q_{\mathcal{H}_0}} D(x) && \text{(C.6)}\\
&= 1 - \sup_{0\leq D\leq 1} \mathbb{E}_{x\sim \hat{q}_{\mathcal{H}_0}} D(x) - \mathbb{E}_{x\sim q_{\mathcal{H}_0}} D(x) && \text{(C.7)}\\
&= \inf_{0\leq D\leq 1} \mathbb{E}_{x\sim q_{\mathcal{H}_0}} D(x) + \mathbb{E}_{x\sim \hat{q}_{\mathcal{H}_0}}(1-D(x)) && \text{(C.8)}\\
&\leq \mathcal{L}_D, && \text{(C.9)}
\end{aligned}$$

for any D in the mentioned space and with equality if and only if $D = D^*$. Eq (5) follows from the Kantorovich-Rubinstein dual representation for general f divergences, proven for example in [183]. □

Corollary 1 *For a generator G with infinite capacity converging to the true conditional distribution $q_{\mathcal{H}_0}(x)$, $\mathcal{L}_G(D^*)$ attains its minimum value of 0.*

Proof. By setting D^* in the loss of the generator $\mathcal{L}_G$ we observe that,

$$\begin{aligned}
\mathcal{L}_G(D^*) &= 1 - \mathcal{L}_{D^*} && \text{(C.10)}\\
&= \frac{1}{2}\int_{\mathcal{X}} |\hat{q}_{\mathcal{H}_0}(x) - q_{\mathcal{H}_0}(x)|dx. && \text{(C.11)}
\end{aligned}$$

Hence, for a generator with infinite capacity converging to the true conditional distribution $q_{\mathcal{H}_0}(x)$, the last term is 0 which implies $\mathcal{L}_G(D^*) = 0$. □

Proof of Theorem 1. Our derivation is similar to [20]. By definition the statistic $\hat{\rho}$ results in a p-value $p < \alpha$ if and only if the observed variable x is contained in the set $A_\alpha := \{x : \sum_{m=1}^{M} \mathbf{1}\{\rho(x^{(m)}, y, z) \geq \rho(x, y, z)\}/M < \alpha\}$. Consider generating a new sample $\tilde{x}$ from the generator G under the estimated conditional distribution and let $x \sim q_{\mathcal{H}_0}$ be sampled from the true conditional. Then it holds that,

$$\begin{aligned}
\mathcal{L}_G(D^*) &= \mathbb{E}_{x\sim\hat{q}_{\mathcal{H}_0}} D^*(x) - \mathbb{E}_{x\sim q_{\mathcal{H}_0}} D^*(x) && \text{(C.12)}\\
&= \int_{\mathcal{X}} |q_{\mathcal{H}_0}(x) - \hat{q}_{\mathcal{H}_0}(x)| dx && \text{(C.13)}\\
&= \sup_A |q_{\mathcal{H}_0}(A) - \hat{q}_{\mathcal{H}_0}(A)| && \text{(C.14)}\\
&\geq Pr(x \in A_\alpha) - Pr(\tilde{x} \in A_\alpha) && \text{(C.15)}\\
&\geq Pr(\hat{\rho} > c_\alpha | \mathcal{H}_0) - \alpha, && \text{(C.16)}
\end{aligned}$$

where by expanding the expectations, eq. (13) follows from similar arguments to those presented in Lemma 1. Next eq. (14) follows from a well known equivalent representation of the total variation divergence between probability measures, proven for example in Proposition 4.2, page 48 of [105]. Eq. (15) follows by standard properties of the supremum operator. Finally we arrive at eq. (16) given the fact that, by definition of A_α, $Pr(x \in A_\alpha) = Pr(\hat{\rho} > c_\alpha | \mathcal{H}_0)$ and, since given y and z the set $(\tilde{x}, x^{(1)}, ..., x^{(M)})$ is conditionally independent and therefore exchangeable, we have that $Pr(\tilde{x} \in A_\alpha) \leq \alpha$. □

Proof of equation (9)**.** Let $q_{\mathcal{H}_0}$ and $q_{\mathcal{H}_1}$ be the true conditional distributions of $X|Z$ under the null hypothesis $H_0 : X \perp\!\!\!\perp Y|Z$ and its alternative $H_1 : X \not\perp\!\!\!\perp Y|Z$ respectively. Denote by A the event that samples x result in a p-value below the level α. Then,

$$\begin{aligned}
\text{Type I error} + \text{Type II error} &= q_{\mathcal{H}_0}(A) + q_{\mathcal{H}_1}(A^c) && \text{(C.17)}\\
&= 1 + q_{\mathcal{H}_0}(A) - q_{\mathcal{H}_1}(A) && \text{(C.18)}\\
&\geq 1 + \inf_A \left(q_{\mathcal{H}_0}(A) - q_{\mathcal{H}_1}(A)\right) && \text{(C.19)}\\
&= 1 - \sup_A \left(q_{\mathcal{H}_1}(A) - q_{\mathcal{H}_0}(A)\right) && \text{(C.20)}\\
&= 1 - \delta_{TV}(q_{\mathcal{H}_0}, q_{\mathcal{H}_1}). && \text{(C.21)}
\end{aligned}$$

□

C.4 Implementation details

GCIT. In all our experiments we have set the depth of the generator, the discriminator and information network to 3. The number of hidden nodes in each layer is $d/10$ and $d/16$ for the generator and discriminator respectively (d the number of inputs). For the information network, we use 2 diagonal matrices for each layer to make two hidden nodes for each feature separately. We use ReLu and tanh as the activation functions for each layer except for the output layer where we use a linear activation function for the information network, and sigmoid activation function for the discriminator and generator network given that we require its output to be constrained in the $(0,1)$ interval and re-scale the data in the $(0,1)$ interval prior to training. The number of samples in each mini-batch is 128 for the synthetic experiments and 64 for the genetic experiment. The GCIT and all experiments have been implemented and carried out in tensorflow and python. Pseudocode for the GCIT is given in Algorithm 1 and a python implementation is given at `https://github.com/alexisbellot/GCIT`.

Baseline algorithms. We implemented the KCIT and RCoT with code provided by the authors in [167] in their R package `RCIT`. The CCIT [156] was implemented with the code provided at `https://github.com/rajatsen91/CCIT/blob/master/CCIT` by the authors. The CRT was implemented in python with our own code.

Algorithm 1 GCIT

Input: batch size n_b, data $\mathcal{D} = (\mathbf{x}, \mathbf{y}, \mathbf{z})$ of size N, statistic ρ, iterations M, parameter λ

Initialize: neural network model parameters ϕ, η, θ

while convergence criteria not satisfied **do**

1. **Update Discriminator**

Sample $\mathbf{z}_1, ..., \mathbf{z}_{n_b}$ from $\mathcal{D}$ and $\mathbf{v}_1, ..., \mathbf{v}_{n_b} \sim p_v$ a batch from the real and latent samples

$\tilde{\mathbf{x}}_i \leftarrow G_\phi(\mathbf{z}_i, \mathbf{v}_i)$ for $i = 1, ..., n_b$

Update η by stochastic gradient descent with,

$$\nabla_\eta \frac{1}{n_b} \sum_{i=1}^{n_b} D_\eta(\mathbf{x}_i, \mathbf{z}_i) + (1 - D_\eta(\tilde{\mathbf{x}}_i, \mathbf{z}_i))$$

2. **Update Information Network**

Sample $\mathbf{z}_1, ..., \mathbf{z}_{n_b}$ from $\mathcal{D}$, $\mathbf{v}_1, ..., \mathbf{v}_{n_b} \sim p_v$ and κ a permutation of $1, ..., n_b$

$\tilde{\mathbf{x}}_i \leftarrow G_\phi(\mathbf{z}_i, \mathbf{v}_i)$ for $i = 1, ..., n_b$

Update θ by stochastic gradient ascent with,

$$\nabla_\theta \left(\frac{1}{n_b} \sum_{i=1}^{n_b} T_\theta(\tilde{\mathbf{x}}_i, \mathbf{x}_i) - \log[\frac{1}{n_b} \sum_{i=1}^{n_b} \exp(T_\theta(\tilde{\mathbf{x}}_i, \mathbf{x}_{\kappa(i)}))] \right)$$

3. **Update Generator**

Sample $\mathbf{z}_1, ..., \mathbf{z}_{n_b}$ from $\mathcal{D}$ and $\mathbf{v}_1, ..., \mathbf{v}_{n_b} \sim p_v$

$\tilde{\mathbf{x}}_i \leftarrow G_\phi(\mathbf{z}_i, \mathbf{v}_i)$ for $i = 1, ..., n_b$

Update ϕ by stochastic gradient descent with,

$$\nabla_\phi (\mathcal{L}_G(D) + \lambda \mathcal{L}_{Info.})$$

end while

for $m = 1, ..., M$ **do**

Sample $v_1, ..., v_N \sim p_v$

$\tilde{x}_j^{(m)} \leftarrow G_\phi(z_j, v_j)$ for $j = 1, ..., N$

$\hat{\rho}^{(m)} \leftarrow \rho(\tilde{\mathbf{x}}^{(m)}, \mathbf{y}, \mathbf{z})$

end for

$\hat{\rho} \leftarrow \rho(\mathbf{x}, \mathbf{y}, \mathbf{z})$

$\hat{p} \leftarrow \sum_{m=1}^{M} \mathbf{1}\{\hat{\rho}^{(m)} \geq \hat{\rho}\}/M$

Output: p-value $\hat{p}$

C.5 Genomics experiment details

The Cancer Cell Line Encyclopedia (CCLE) is a compilation of gene expression, chromosomal copy number and sequencing data from 947 human cancer cell lines. A cancer cell line can be understood as a string of cancer cells that keep dividing and growing over time under certain conditions in a laboratory. Then, using high-throughput sequencing technologies, the molecular characteristics of cancer cell lines, such as gene expression or mutation data, can be extracted. These genetic predictors were coupled with measures of drug sensitivity for *PLX*4720: a drug used against cancer whose response is available for 474 of the above cancer cell lines. By correlating the genetic information with the corresponding sensitivity to drug response, the data in principle allows for the identification of relevant genetic markers which could then lead to personalized treatment therapies depending on a patients genetic makeup. We illustrate this procedure in Figure C.5.

Except for the conditional independence test to report significant genetic variables, our experiments followed similar procedures to those detailed in [173] and [9]. We choose to analyze dependence of drug response with 466 genetic mutations observed on each cancer line. We give summary statistics of the final data used in Table C.1 below. This is a very high-dimensional problem that makes conditional independence testing unfeasible with traditional tests.

As in the original study in [9], we proceeded by fitting an elastic net model to predict drug response from genetic features with 10-fold cross-validation to optimize hyperparameters. Influential features were then ranked by their heuristic importance score given by the magnitude of fitted parameter values. The random forest model was used with default hyper-parameters in the python library `sklearn` and the CRT was implemented with a Gaussian approximation like in all other experiments. We ran the GCIT and the CRT considering each feature separately with drug response and all remaining features as confounders.

Remark. For a more systematic biological evaluation of features reported by the GCIT, we would use a more principled feature selection procedure such as Benjamini-Hochberg's correction for false discoveries [16].

Table C.1: Summary statistics of the final genetic data used from [9].

Statistics	**Values**
No. of cancer cell lines	474
No. of genetic mutations	466
Pearson correlation with drug response	**min:** 0.05, **max:** 0.51, **mean:** 0.07, **var:** 0.001
Drug response distribution	**min:** -97.9, **max:** 43.3, **mean:** -17.2, **var:** 633.3

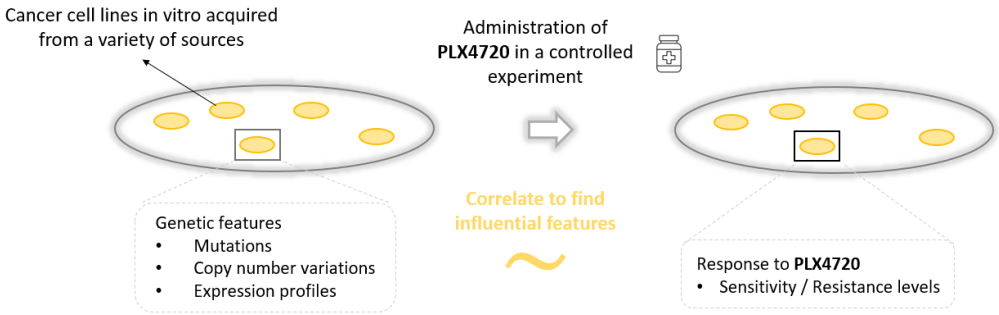


Figure C.5: Diagram illustrating the data used in the Genetic experiment.

Appendix D

Appendix to Chapter 5

This appendix provides additional material accompanying Chapter 5: "Accounting for Unobserved Confounding in Domain Generalization". It is outlined as follows:

- In section D.1 we present additional experiments designed to demonstrate the causal interpretation one may give to DIRM if all conditions for causality are satisfied (i.e. are available interventions on all observed variables). We also include a sensitivity analysis to show the impact on performance of changing the regularization parameter λ.
- In section D.2 we show how measurement error may be interpreted as an instance of unosebrved confounding.
- In section D.3 we provide proofs for the statements made in the main body of this chapter.
- In section D.4 we give additional experimental details, including on the implementation of DIRM and on the datasets used.

D.1 Additional experiments

So far, we have considered predictive performance under different data distributions with selected hyper-parameter configurations of all algorithms to illustrate heterogeneous behaviour of algorithms trained with different learning principles in the presence of unobserved confounders. In this section we revisit our introductory example to investigate in more details learned prediction rules and any sensitivities of interest, especially to hyper-parameter configurations.

We will use the same data generating mechanism presented in the introductory example in Figure 5.1 of the main body of this chapter. Recall that we assume access to observations of variables (X_1, X_2, Y) in two training datasets, each dataset sampled with differing interventions on (X_1, X_2) (in this case differing variances $\sigma^2 = 1$ and $\sigma^2 = 2$) from the following structural model,

$$X_2 := -H + E_{X_2}, \quad Y := X_2 + 3H + E_Y, \quad X_1 := Y + X_2 + E_{X_1} \quad H := E_H,$$

where $E_{X_1}, E_{X_2} \sim \mathcal{N}(0, \sigma^2)$, $E_Y \sim \mathcal{N}(0,1)$, $E_H \sim \mathcal{N}(0,1)$ are exogenous variables. H is an unobserved confounder, not observed during training but that influences the observed association between X_2 and Y.

D.1.1 Recovery of causal coefficients

In this section, given the above two training datasets, we inspect the weights learned in a simple one layer feed-forward neural network to determine exactly whether unobserved confounding induces a given learning paradigm to exploit spurious correlations and to what extent.

By way of preface, we have mentioned that causal, in contrast with spurious, solutions to a prediction problem may be defined as the argument solving,

$$\underset{f}{\text{minimize}} \sup_{P \in \mathcal{P}} \mathbb{E}_{(x,y)\sim P}[\mathcal{L}(f(x), y)], \tag{D.1}$$

for $\mathcal{P}$ defined as any distribution arising from *arbitrary* interventions on observed covariates x leading to shifts in their distribution P_x (see sections 3.2 and 3.3 in [117] for a detailed discussion of this result). This objective is a special case of the proposed optimization problem (5.5), specifically it is an affine combination (with $\lambda \to \infty$) of distributions with different shifts in P_x in all observed variables x.

We demonstrate this fact empirically in Table D.1. In principle, causal solutions are recoverable with the proposed approach because we do observe during training environments with shifts in $p(X_1, X_2)$, irrespective of the presence or not of unobserved confounders. We see that this holds approximately for the proposed objective with estimated coefficients $(0.01, 0.95)$ for (X_1, X_2) close to the true causal coefficients $(0, 1)$. In contrast, ERM returns biased coefficients and so does IRM.

This empirical observation is important because it highlights the fact that enforcing minimum gradients on average (ERM) or simultaneously across environments (the regularization pro-

posed by IRM) is not appropriate to recover causal coefficients in the presence of unobserved confounders.

If however, no unobserved confounders exist in the system being modelled ($H := 0$ in the data generating mechanism) our objective and IRM are equivalent in the limit, and estimated parameters coincide with the causal solution approximately. This experiment is given in Table D.2.

Table D.1: Bias in estimation **with** unobserved confounders.

	Truth	**ERM**	**IRM** ($\lambda \to \infty$)	**DIRM** ($\lambda \to \infty$)
Estimated parameters	[0, 1]	[0.91, -1.02]	[0.75, -0.76]	[0.01, 0.95]

Table D.2: Bias in estimation **without** unobserved confounders.

	Truth	**ERM**	**IRM** ($\lambda \to \infty$)	**DIRM** ($\lambda \to \infty$)
Estimated parameters	[0, 1]	[0.5, -0.6]	[0.01, 0.98]	[0.02, 0.96]

D.1.2 Sensitivity to hyper-parameters

The robustness guarantees of any particular solution depends on the extent of the extrapolation desired (as a function of λ). For larger values of this parameter we can expect solutions to be robust in a larger set of distributions, spanning empirical risk minimization for $\lambda = 0$, to convex combinations, to training environments to arbitrary affine combinations of training environments for increasing λ.

In this section, we analysed performance in test data with the exact same data generating mechanism as considered in the introduction of the main body of this chapter as a function of λ. Figure D.1 gives our performance results that empirically verifies that the proposed approach interpolates between empirical risk minimization and causality in this case. We can see that for λ approaching zero solutions converge to ERM, for $\lambda = 2$ the solutions was equivalent to DRO, and for increasing λ the solutions approximate the causal one in the limit.

D.2 Other examples of unobserved confounding

Measurement error. The data generating processes described in the main body of this chapter for instance, as well as most of machine learning, assume that all nuisance variability enters

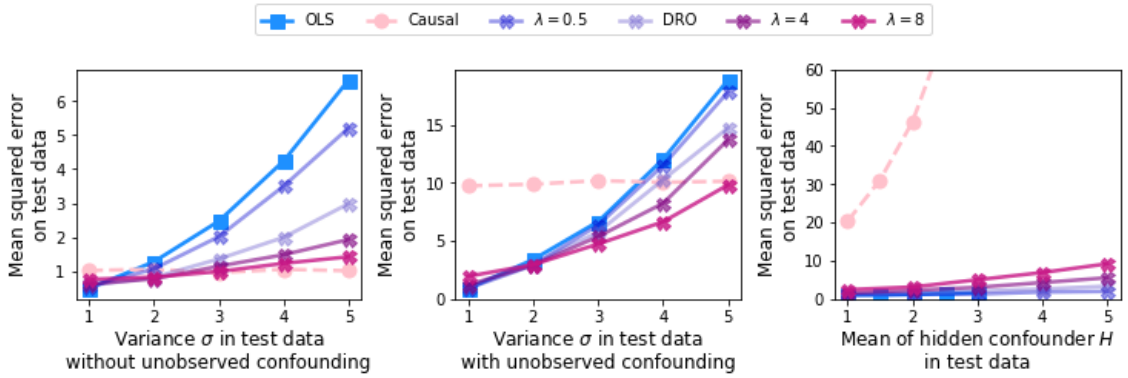


Figure D.1: **Sensitivity of solutions to hyperparameter** λ. Ordinary Least Squares (OLS) and the causal solution, with coefficients $(0,1)$ for (X_1, X_2) are two extremes ($\lambda = 0$ and $\lambda \to \infty$ respectively) of the spectrum of solutions that can be attained with the proposed approach. Positive values of λ interpolate in some sense between OLS and causal solutions in this case. Here DRO corresponds to DIRM with $\lambda = 2$, approximately.

the causal mechanisms of the data; that is, observed data reflects *only* causal drivers. If this is not the case, for example because of independent measurement noise observed in data but that does not propagate to across causal children, regression is known to be inconsistent in general [32] and its bias is analogous to a form of unobserved confounding.

Consider a simple model for illustration. Suppose (X, Y) are observed subject to measurement noise, $X^\star = X + E_x$ and $Y^\star = Y + E_y$, which are not causally related to one another but rather $Y = \beta X + E$. Let $E_x = \beta_x H$ and $E_y = \beta_y H$ be the structure of measurement error independent of X and Y. Then substituting our observed data $(X^\star, Y^\star)$ into the underlying (X, Y) relationship the observed model is,

$$Y^\star = \beta X^\star + (\beta_y - \beta_x \beta) H + E, \quad X^\star = \beta_x H + X. \tag{D.2}$$

A special case of regression with unobserved confounders H.

D.3 Technical results

This section provides a more complete discussion of the assumptions and justification statements relating to causality in section 5.1.2, and the proof of Theorem 1.

D.3.1 Invariances in the presence of unobserved confounding

In section 5.1.2 we justified exploiting a certain invariance of causal coefficients in the inner product of functions of the data X and residuals E, to occur even in the presence of unobserved confounders as long as interventions that define different environments do not involve unobserved confounders H.

Here we show this invariance to hold in the special case of an additive model. The general data generation mechanism is as follows. Data sources, or different environments, emerge from manipulations in exogenous E_X, related to X only, in an underlying additive model $\mathbb{F}$ with also additive functions f_1, f_2, f_3, f_4,

$$Y := f_1(X) + f_2(H) + E_Y, \qquad X := f_3(X) + f_4(H) + E_X, \qquad H := E_H. \tag{D.3}$$

Exogenous variables (E_X, E_Y, E_H) may have arbitrary distributions but only E_X or E_Y vary across environments. Then it holds that,

$$\begin{aligned} X &= (I - f_3)^{-1} f_4(H) + (I - f_3)^{-1} E_X \\ &= (I - f_3)^{-1} f_4(E_H) + (I - f_3)^{-1} E_X, \end{aligned}$$

and that,

$$\nabla_\beta f_1(X)(Y - f_1(X)) = \left(\nabla_\beta f_1 (I - f_3)^{-1} f_4(E_H) + \nabla_\beta f_1 (I - f_3)^{-1} E_X\right) \cdot (f_2(E_H) + E_Y),$$

which is a product of functions involving E_H in one term, E_H and E_Y in another term, E_X and E_H in another term, and E_X and E_Y in the last term. Since (E_X, E_Y, E_H) are mutually independent taking expectations of product of functions involving E_X and E_H, E_X and E_H, and, E_X and E_Y equals 0 assuming $f_i(E_j) = 0$ for $i = 1, \ldots, 4$ and $j \in \{X, Y, H\}$.

So concluding, the expectation of the inner product $\nabla_\beta f_1(X)(Y - f_1(X))$ does not depend on E_X nor E_Y and is thus stable across environments that have changing distributions for E_X or E_Y. Now note that other functions than f_1 may have this property as well, i.e. predictors that satisfy this invariance are not necessarily unique and will depend on the differences between available environments. If however, only one predictor exist that satisfies this invariance we may say that this predictor is causal. We summarize this claim in the following statement.

Proposition 1 *Let Y and X be related by a non-linear additive model with unobserved confounding as in (D.3). Then,*

$$\mathbb{E}_{P_i}\nabla_\beta f_1(X)(Y - f_1(X)) = \mathbb{E}_{P_j}\nabla_\beta f_1(X)(Y - f_1(X)), \tag{D.4}$$

under the assumption that distributions on (X,Y) P_i and P_j are given by a data generating mechanism (D.3) subject to interventions on E_X or E_Y only. Moreover, a function f satisfying the above equality, if unique is equal to f_1.

D.3.2 Proof of Theorem 1

We restate the Theorem for convenience.

Theorem 1 *Let $\{P_e\}_{e\in\mathcal{E}}$, be a set of available environments. Further let the parameter space of β be open and bounded, such that the expected loss function $\mathcal{L}$ as a function of β belongs to a Sobolev space. Then, the following inequality holds,*

$$\begin{aligned}
&\sup_{\alpha_e\in\Delta_\eta} \sum_{e\in\mathcal{E}} \alpha_e \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x),y\right) \leq \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f\circ\phi(x),y\right) \\
&+ (1+n\eta)\cdot C\cdot \left|\left| \sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \nabla_\beta \mathcal{L}\left(f\circ\phi(x),y\right) - \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \nabla_\beta \mathcal{L}\left(f\circ\phi(x),y\right) \right|\right|_{L_2},
\end{aligned}$$

where C depends on the domain of β, $n := |\mathcal{E}|$ is the number of available environments and $e\sim\mathcal{E}$ loosely denotes sampling indeces with equal probability from $\mathcal{E}$.

Proof. Let Ω denote the parameter space of β. The following derivation shows the claim,

$$\begin{aligned}
&\sup_{\alpha_e\in\Delta_\eta} \sum_{e\in\mathcal{E}} \alpha_e \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x),y\right) \\
&= (1+n\eta)\cdot \sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x),y\right) - \eta \sum_{e\sim\mathcal{E}} \mathbb{E}_{P_e} \mathcal{L}\left(f\circ\phi(x),y\right) \\
&= \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f\circ\phi(x),y\right) + (1+n\eta)\cdot \sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x),y\right) \\
&- (\eta+1/n) \sum_{e\sim\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x),y\right) \\
&= \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f\circ\phi(x),y\right) \\
&+ (1+n\eta)\cdot \left(\sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x),y\right) - \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f\circ\phi(x),y\right) \right) \\
&\leq \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f\circ\phi(x),y\right) \\
&+ (1+n\eta)\cdot M\cdot \left|\left| \sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x),y\right) - \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f\circ\phi(x),y\right) \right|\right|_{L_2},
\end{aligned}$$

where the inequality is given by the property that the evaluation functional is a bounded linear operator in certain Sobolev spaces $\mathcal{W}$, for example with $\Omega = \mathbb{R}^d$ and L_2 norm. In particular

this means that $|f(\beta)| \leq M||f||_{L_2}$ for all $f \in \mathcal{W}$. It follows then also that the above is,

$$\leq \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}(f\circ\phi(x),y) +$$
$$(1+n\eta)\cdot P\cdot M\cdot \left|\left| \sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \nabla_\beta \mathcal{L}(f\circ\phi(x),y) - \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \nabla_\beta \mathcal{L}(f\circ\phi(x),y) \right|\right|_{L_2},$$

by Poincaré's inequality for Sobolev functions defined on an open, bounded parameter space, see e.g. [104]. The assumption we make here for this last inequality to hold is that the region where the difference in loss functions is near zero is large enough such that the integral of the gradient is also large enough to control the integral of the function. This holds however for functions defined on many "reasonable" parameter spaces (Lipschitz suffices).

D.4 Experimental details

This section gives implementation details of DIRM, additional experiments to test sensitivities relating to optimization choices, and a complete description of the data and experiments performed in the main body of this chapter.

D.4.1 Implementation details

The regularizer in DIRM's objective in equation (5.5) controls the regularity or variation of the prediction function and encourages to learn a representation ϕ that results in a prediction function with similar derivatives in all training domains. The L_2 norm integrates out the influence of β in the regularizer and thus most of the optimization involves ϕ, though β still plays a role in the first term of the objective.

In all our experiments, $f : \mathbb{R}^h \to \mathbb{R}$ as well as $\phi : \mathbb{R}^d \to \mathbb{R}^h$ are implemented as fully connected neural networks (ϕ with optional hidden layers). β can thus be interpreted as the weights and biases of f. The L_2 norm must be approximated in practice, which we do by evaluating the vector norm of the derivative of f with respect to β on a batch of training examples of each environment. The variance on the computed norms between environments is a proxy for the maximum deviation between environments with a smoother gradient vector field. Each step of the optimization then alternates between an update on ϕ and update on f, as detailed in the algorithm below.

DIRM is sensitive to initialization and to the choice of hyperparameters – specifically its optimization schedule. In our experiments, we found best performance by increasing the relative weight of the penalty term λ after a fixed number of iterations (and similar implementations are used for IRM and REx that suffer from similar challenges). This we believe could be a significant limitation for its use in practice since this choice must be made a priori. We investigated the sensitivity of DIRM to this optimization schedule in Table D.3 that shows test accuracy as a function of the iteration at which penalty term weight λ is increased.

Table D.3: Test set performance (accuracy in %) on X-ray data as a function of the number of epochs used to increase penalty λ.

	2 epochs	**4 epochs**	**6 epochs**	**8 epochs**	**10 epochs**	**12 epochs**
IRM	56.4 ($\pm$ 6)	58.1 ($\pm$ 3)	59.2 ($\pm$ 3)	59.1 ($\pm$ 2)	58.1 ($\pm$ 3)	57.8 ($\pm$ 1)
REx	55.3 ($\pm$ 8)	57.9 ($\pm$ 4)	60.6 ($\pm$ 3)	60.5 ($\pm$ 2)	57.7 ($\pm$ 3)	54.7 ($\pm$ 2)
DIRM	54.1 ($\pm$ 7)	61.7 ($\pm$ 4)	63.8 ($\pm$ 3)	63.5 ($\pm$ 3)	62.6 ($\pm$ 2)	58.2 ($\pm$ 1)

Choosing this number accurately is important for generalization performance. If λ is increased too early, different initialization values (and the complex loss landscape) lead to different solutions with unreliable performance and a large variance. This happens for all methods. An initial number of iterations minimizing loss in-sample improves estimates for all methods which then converge to solutions that exhibit lower variance.

Algorithm 2 DIRM

Input: datasets $\mathcal{D}_1, \ldots, \mathcal{D}_E$ in E different environments, parameter λ, batch size K

Initialize: neural network model parameters ϕ, β

while convergence criteria not satisfied **do**

for $e = 1, \ldots, E$ **do**

Estimate loss $\mathcal{L}_e(\phi, \beta)$ empirically using a batch of K examples from $\mathcal{D}_e$.

Estimate derivatives $\nabla_\beta \mathcal{L}_e(\phi, \beta)$ empirically using a batch of K examples from $\mathcal{D}_e$.

end for

Update β by stochastic gradient descent with,

$$\nabla_\beta \left(\frac{1}{E} \sum_{e=1}^{E} \mathcal{L}_e(\phi, \beta) \right)$$

Update ϕ by stochastic gradient descent with,

$$\nabla_\phi \left(\frac{1}{E} \sum_{e=1}^{E} \mathcal{L}_e(\phi, \beta) + \lambda \cdot \mathrm{Var}(||\nabla_\beta \mathcal{L}_1(\phi, \beta)||_2^2, \ldots, ||\nabla_\beta \mathcal{L}_E(\phi, \beta)||_2^2) \right)$$

end while

Approximation of the L_2 norm in practice

The bound given in Theorem 1 quantifies the discrepancy between function derivatives using the L_2 norm, defined as an intregal over possible parameter values β. For neural networks, computation of the L_2 norm is largely intractable and specifically, for networks of depth greater or equal to 4, it is an NP-hard problem (see Proposition 1 in [177]). Some approximation is thus unavoidable. One option is to recognise the L_2 norm as an expectation over functional evaluations, $||f||_{L_2} = \mathbb{E}_{x \sim \mathcal{U}(\Theta)} \left[||f(x)||_2^2 \right]^{1/2}$ for a continuous function f taking values x sampled uniformly from its domain Θ. Empirical means are tractable yet they induce a much higher computational burden as these must be computed in every step of the optimization. Our approach is to take this approximation to its limit, making a single function evaluation at each step of the optimization using the current estimate β, as written in Algorithm 1.

This approximation loosens the connection between the bound given in Theorem 1 and the proposed algorithm. It remains justified however from a conceptual perspective as the objective of controlling an L_2 type of norm is to encourage the regularizer function towards 0, and thus the values of the regularizer (which we do explicitly). For empirical comparisons with the empirical mean approach, we implemented empirical means using all combinations of parameter values chosen from a grid of 5 parameter values around the current estimate β, $\{0.25\beta, 0.5\beta, \beta, 2\beta, 4\beta\}$. Table D.4 shows similar performance across the real data experiments considered in the main body of this chapter. A single evaluation is in practice enough to monitor invariance of representations to environment-specific loss derivatives.

Table D.4: Test set performance (accuracy in %) on real datasets for two different regularizer approximations to the L_2 norm.

	Pneumonia Prediction	**Parkinson Prediction**	**Survival Prediction**
DIRM-means	63.5 ($\pm$ 3)	73.0 ($\pm$ 1.5)	78.0 ($\pm$ .9)
DIRM-single	63.7 ($\pm$ 3)	72.8 ($\pm$ 2)	77.9 ($\pm$ 1)

D.4.2 Data details

X-ray data

We create training environments with different proportions of X-rays from our two hospital sources to induce a correlation between the hospital (and its specific data collection procedure) and the pneumonia label. The objective is to encourage learning principles to exploit a spurious correlation, data collection mechanisms should not be related to the probability of being diagnosed with pneumonia. The reason for creating two training data sets with slightly different spurious correlation patterns is to nevertheless leave a statistical footprint in the distributions to disentangle stable (likely causal) and unstable (likely spurious). In each of the training and testing datasets we ensured positive and negative labels remained balanced. The training datasets contained 2002 samples each and the testing dataset contained 1144 samples.

All learning paradigms trained a convolutional neural network, 2 layers deep, with each layer consisting of a convolution (kernel size of 3 and a stride of 1). All predictions were made off of the deepest layer of the network. The number of input channels was 64, doubled for each subsequent layer, and dropout was applied after each layer. We optimize the binary cross-entropy loss using Adam (learning rate 0.001) without further regularization on parameters and use Xavier initialization. While learning with IRM and the proposed approach, the respective penalty $\lambda = 1$ is added to the loss after 5 epochs of learning with $\lambda = 0$. Experiments are run for a maximum of 20 epochs with early stopping based on validation performance. All results are averaged over 10 trials with different random splits of the data, and the reported uncertainty intervals are standard deviations of these 10 performance results.

Parkinson's disease speech data

The data includes a total of 26 features recorded on each sample of speech and set training and testing splits which we use in our experiments. For each patient 26 different voice samples including sustained vowels, numbers, words and short sentences where recorded, which we considered to be different but related data sources. We created three training environments by concatenating features from three number recordings, concatenating features from three word recording and concatenating features from three sentences; for a total of 120 samples in each of the three training environment. The available testing split contained 168 recordings of vowels, which we expect to differ from training environments because these are different patients and

do not contain numbers or words. Positive and negative samples were balanced in both training and testing environments.

On this data, for all learning paradigms we train a multi-layer perceptron with two hidden layers of size 64 with tanh activations and dropout ($p = 0.5$) after each layer. As in the image experiments, we optimize the binary cross-entropy loss using Adam (learning rate 0.001), L_2 regularization on parameters and use Xavier initialization. While learning with IRM and the proposed approach, the respective penalty is added to the loss $\lambda = 1$ after 200 epochs of learning with $\lambda = 0$ to ensure stable optimization. Experiments are run for a maximum of 1000 epochs with early stopping based on the validation performance. All results are averaged over 10 trials with different random seeds of our algorithm. This is to give a sense of algorithm stability rather than performance stability.

MAGGIC electronic health records

MAGGIC stands for Meta-Analysis Global Group in Chronic Heart Failure. The MAGGIC meta-analysis includes individual data on $39,372$ patients with heart failure (both reduced and preserved left-ventricular ejection fraction), from 30 cohort studies, six of which were clinical trials. 40.2% of patients died during a median follow-up of 2.5 years. For our purposes, we removed patients that were censored or lost to follow-up to ensure well-defined outcomes after 3 years after being discharged from their respective hospitals. A total of 33 variables describe each patient including demographic variables: age, gender, race, etc; biomarkers: blood pressure, haemoglobin levels, smoking status, ejection fraction, etc; and details of their medical history: diabetes, stroke, angina, etc.

To curate our training and testing datasets, we proceeded as follows. On all patients follow-up over 3 years, we estimated feature influence of survival status after three years. A number of variables were significantly associated with survival out of which we chose Age, also found correlated with BMI and a number of medical history features, as a confounder for the effect of these variables on survival. We used three criteria to select studies: having more than 500 patients enrolled and balanced death rates (circa 50%). 5 studies fitted these constraints: 'DIAMO', 'ECHOS', 'HOLA', 'Richa', 'Tribo'. Each was chosen in turn as a target environment with models trained on the other 4 training environments.

Feature reproducibility experiments. A natural objective for the consistency of health care and such that we may reproduce the experiments and their results in different scenarios is to find relevant features that are not specific to an individual medical study, but can also be

found (replicated) on other studies with different patients. Heterogeneous patients and studies, along with different national guidelines and standards of care make this challenging. In our experiments we made comparisons of reproducibility in parameter estimates for models trained using Empirical Risk Minimization (ERM) and DIRM. We chose networks with a single layer with logistic activation and focused on the estimation of parameter to understand the variability in training among different data sources. Naturally, feature importance measured by parameter magnitudes makes sense only after normalization of the covariates to the same (empirical) variance (equal to 1) in each study separately. After this preprocessing step, for both ERM and the proposed approach we trained separate networks on 100 random pairs of studies (each pair concatenated for ERM) and returned the top 10 significant features (by the magnitude of parameters). Over all sets of significant parameters we then identified how many intersected across a fixed number of the 100 runs.

The same architecture and hyperparameters as in Parkinson's disease speech data experiments was used for MAGGIC data except that we increase the maximum training epochs to 5000.

Appendix E

Appendix to Chapter 6

This appendix provides additional material accompanying Chapter 6: "Scoring of DAGs with Dense Unobserved Confounding". It is outlined as follows:

- Section E.1 proves Theorem 1.
- Section E.2 includes further simulations and details of the synthetic experiments and implementations.
  - Section E.2.1 gives details of the synthetic experiments, metrics of evaluation.
  - Section E.2.2 includes an experiment analysing performance with sparse unobserved confounding.
  - Section E.2.3 analyses the recovery of the exact weighted adjacency matrix with synthetic simulations.
  - Section E.2.4 gives further reproducibility experiments on skeleton recovery.
- Section E.3 gives details of the (semi-synthetic) genetic experiments.

E.1 Proof of Theorem 1

We begin by recalling the adjusted regression model that we seek to analyse.

$$FX = FX(W+C) + F\bar{E} \qquad \Rightarrow \qquad \tilde{X} = \tilde{X}(W+C) + \tilde{E}. \tag{E.1}$$

Let us write $\tilde{\Sigma} = \text{Cov}(\tilde{X})$ for the covariance matrix of $\tilde{X}$. Even for a good choice of F that balances between a well behaved error term $\tilde{E} = F\bar{E}$, well behaved design matrix $\tilde{X}$ and well behaved perturbation term $\tilde{X}C$ tending to zero, W is not necessarily uniquely identifiable. The map between the observed covariance $\tilde{\Sigma}$ and the pair of causal adjacency matrix W and error covariance $\tilde{\Sigma}_E = \text{Cov}(\tilde{E})$ is not necessarily unique. To avoid issues of identifiability, recent work [4] defines minimum-trace DAGs W_{min},

$$(W_{\text{min}}, \Sigma_{\text{min}}) \in \arg\min\{Tr(\tilde{\Sigma}) : (W, \tilde{\Sigma}_E) \in \mathcal{D}\}, \tag{E.2}$$

where $\mathcal{D}$ denotes all pairs $(W, \tilde{\Sigma}_E)$ that exhibit a data covariance indistinguishable from that observed. Minimum-trace DAGs themselves are not necessarily unique in general but for the purposes of the results presented here we will assume it to be unique for good choices of F that shrink the spurious signal without altering the causal signal too much. We note that extensions exist for unidentifiable case [3], in which case penalized score optimization can be shown to converge to a sparse representative within the class of minimum-trace DAGs but leave this investigation in the presence of unobserved confounding to future work.

Our objective is to control the likelihood of the following failure event,

$$\{\text{supp}(W_{\text{min}}) \subsetneq \text{supp}(\widehat{W})\}, \tag{E.3}$$

where $\widehat{W}$ is the solution to the constrained, penalized optimization program,

$$\widehat{W} \in \underset{W \in \mathbb{D}}{\text{argmin}} \;\; \mathcal{S}(W; \mathbf{X}), \qquad \mathcal{S}(W; \mathbf{X}) := \frac{1}{2n}||\tilde{\mathbf{X}} - \tilde{\mathbf{X}}W||_F^2 + \lambda||W||_1. \tag{E.4}$$

This can be done by reducing the analysis of $\widehat{W}$ to a family of neighbourhood regression problems [4, 3]. There are two key steps:

• First showing that $\widehat{W}$ is equivalent to solving a series of p regression problems given by,

$$\underset{\mathbf{w}_i \in \mathbb{R}^p,\ \text{supp}(\mathbf{w}_i) \subset S}{\arg\min} \;\; \frac{1}{2n}||\tilde{\mathbf{X}}_i - \tilde{\mathbf{X}}\mathbf{w}_i||_2^2 + \lambda||\mathbf{w}_i||_1, \tag{E.5}$$

as defined in the main body of this paper.

• And second, controlling for the error in estimation in each of these neighbourhood problems for all subsets of covariates, or neighbourhoods given by S.

Point 1. The first step is a consequence of how the least squares loss and regularizer factor. This allows to formally establish the equivalence between the DAG problem and neighbourhood regression, and is justified by Lemma B.1. in [3]. This is similar to undirected models, for which the analysis can be reduced to p different regression problems, namely the regression of X_j onto X_{-j}. Unfortunately, for DAGs, there are $p2^p$ possible regression problems (the regression of X_j onto any subset of other variables S), which quickly become intractable to control uniformly. In the identifiable case, we can constrain ourselves to control over sets S that are consistent with a superstructure G of the underlying graph, i.e. we must only control over those adjacency matrices that are sub-graphs of G (e.g. the moral graph of a DAG is an example of superstructure). [3] then show a uniform concentration bound for the score function restricted to a consistent superstructure and use this result to show that any estimated $\hat{W}$ has the same topological sort as $W_{\min}$. This topological sort identifies candidate parent sets for each node X_j, and reduces the problem to control over p regression problems, which is substantially lower than $p2^p$ problems.

These steps rely on the model distribution, independence of the error term in (E.1), and the properties of minimum-trace DAGs, and are given as a sequence of Lemmas and Propositions in Appendix B in [3]. All proofs (and prior conditions for the applicability of each statement) therein hold for our model without modification since the distribution family is preserved under deterministic transformations of both sides of the model equation, and the independence of error terms holds by construction of the matrix C and Gaussianity. We refer the reader to these references for a detailed derivation of each of these steps.

Point 2. The second point differs from [3]. It holds that the optimization program (E.4) can be reduced to a collection of local regression problems, but in our case each regression problem is defined as (E.5) rather than the conventional un-adjusted lasso. For this problem, as mentioned, a good choice of F needs to find a balance between a well behaved error term $\tilde{E} = F\bar{E}$, well behaved design matrix $\tilde{X}$ and well behaved perturbation term $\tilde{X}C$. These conditions can be articulated in three assumptions on the adjusted program.

- We assume $\lambda_{\max}(\mathrm{Cov}(X,H)) = \mathcal{O}(\sqrt{p})$: the largest singular value of the $(p \times q)$ covariance matrix of (X,H) is of the order $\sqrt{p}$, which is a consequence of denseness of unobserved confounding (the effect on each individual observable being small but spread over a large number of variables).
- We assume that $\tilde{d}_{n/2} = \mathcal{O}(\sqrt{p})$: the median value of the singular values of $\tilde{X}$ is of the order $\sqrt{p}$, with high probability.

- We assume that the compatibility constant ϕ_M of $M := n^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$ is of the same order as the minimal singular value of X. The compatibility constant is a kind of restricted eigenvalue condition and is common in the model selection literature. For a square matrix M it is defined as,

$$\phi_M := \inf_{||\alpha||_1 \leq 5||\alpha_S||_1} \frac{\sqrt{\alpha^T M \alpha}}{||\alpha_S||_1/\sqrt{s}}, \tag{E.6}$$

  where S is the support set of $\mathbf{w}_i$, s is the size of S and α_S is a vector consisting only of the components of α which are in S.

With these conditions, [33] demonstrated that the error in estimation of $\mathbf{w}_i$ with the program (E.5) to be bounded in l_1 norm by a factor of order,

$$\mathcal{O}\left(\frac{\sigma_i s}{\sigma_{min}(\tilde{\Sigma}_E)}\sqrt{\log p/n}\right), \tag{E.7}$$

where $\sigma_{min}(M)$ denotes the smallest singular value of a matrix M, σ_i is the standard deviation of $\tilde{E}_i$ and s is the size of the support of $\mathbf{w}_i$.

Control over events of the form $\{\text{supp}(\mathbf{w}_i) \subsetneq \text{supp}(\hat{\mathbf{w}}_i)\}$ then follows with an additional beta-min condition, i.e. a condition minimum strength on the signal of causal coefficients,

$$\min(|w| : w \in \text{supp}(W_{\min})) \gtrsim \sigma\sqrt{\log p/n},$$

where we have written $a \gtrsim b$ to mean that $a \geq C \cdot b$ for some constant $C > 0$, and $\sigma = \frac{\max_i(\sigma_i)s}{\sigma_{min}(\tilde{\Sigma}_E)}$. To see this notice that,

$$||\hat{\mathbf{w}}_i - \mathbf{w}_i||_1 \geq ||\hat{\mathbf{w}}_i - \mathbf{w}_i||_\infty. \tag{E.8}$$

It follows that $\text{supp}(\mathbf{w}_i) \subseteq \text{supp}(\hat{\mathbf{w}}_i)$ as long as $\min(|w| : w \in \text{supp}(\mathbf{w}_i)) \gtrsim \sigma\sqrt{\log p/n}$ with high probability. If not, we could find a $j \in \text{supp}(\mathbf{w}_i)$ with $j \notin \text{supp}(\hat{\mathbf{w}}_i)$ such that $|\hat{w}_{ij} - w_{ij}| = |\hat{w}_{ij}| \gtrsim \sigma\sqrt{\log p/n}$, which leads to a contradiction. Here w_{ij} is the j-th element of the vector $\mathbf{w}_i$.

Finally, control over false positives $\{\text{supp}(\mathbf{w}_i) \nsubseteq \text{supp}(\hat{\mathbf{w}}_i)\}$ in each neighbourhood regression problem implies control over events $\{\text{supp}(W_{\min}) \nsubseteq \text{supp}(\widehat{W})\}$ in DAG estimation by a uniform bound over the control ensured in the p distinct neighbourhood regression problems, and is technically justified by point (b) in Lemma B.1 in [3], that ensures that $\widehat{W}$ is the unique solution to (E.4) if and only if $\hat{\mathbf{w}}_\mathbf{i} = [\widehat{W}]_{\cdot i}$ is the unique solution to (E.5).

E.2 Details on synthetic experiments

E.2.1 Simulations, metrics and implementation

In the main body of this paper, we consider one main synthetic network model:

- Erdös–Rényi graph models. These are generated by adding edges independently with equal probability $r = \frac{2e}{p^2-p}$, where e is the expected number of edges in the resulting graph. For each p-node graph, we simulate graphs with e equal to p.

Based on the DAG sampled from this graph model, we assign edge weights sampled independently from Uniform$([-2,-0.5] \cup [0.5,2])$ to construct the weighted adjacency matrix $W \in \mathbb{R}^{d\times d}$. The observational data is then generated according to the linear confounded DAG model with different graph sizes, and additive noise types:

- Gaussian. $H_i, E_j \sim \mathcal{N}(0,1), \quad i = 1,\ldots,q, \quad j = 1,\ldots,p.$
- Exponential. $H_i, E_j \sim \text{Exp}(1), \quad i = 1,\ldots,q, \quad j = 1,\ldots,p.$
- Gumbel. $H_i, E_j \sim \text{Gumbel}(0,1), \quad i = 1,\ldots,q, \quad j = 1,\ldots,p.$

In each synthetic experiment we generate $n = 100$ samples for each of these settings. For experiments considering performance as a function of varying dimensionality p of X, we fixed $q = 10$ and $\sigma = 0.2$. For experiments considering varying dimensionality q of H, we fixed $p = 20$ and $\sigma = 0.2$. For experiments considering varying σ, we fixed $p = 20$ and $q = 10$.

We evaluate the estimated graphs using four different metrics:

- Structural Hamming Distance (SHD) indicates the number of edge additions, deletions, and reversals in order to transform the estimated graph into the ground truth DAG.
- True Positive Rate (TPR) measures the proportion of actual positive edges that are correctly identified as such.
- False Discovery Rate (FDR) measures the proportion of false discoveries among the estimated edges.
- The Area Under the ROC Curve (AUC) measures the area under a plot of the TPR as a function of FDR as te threshold for determining presence / absence of edges is varied.
- The l_2 loss in the recovery of adjacency matrices $||\widehat{W} - W||_2^2/p$.

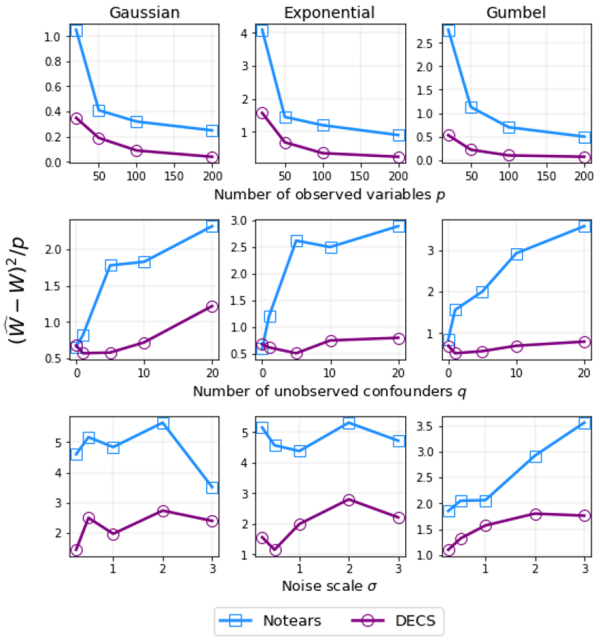


Figure E.1: Performance on the recovery of the weighted adjacency matrix.

We use the following implementations for baseline algorithms.

- FCI was implemented through the `pcalg` R package with a Gaussian conditional independence test.
- LGES was implemented with hyperparameters chosen by cross validation following the author's implementation at `https://github.com/benjaminfrot/lrpsadmm/`.
- NOTEARS. We use the variant with l_1 regularization chosen by cross-validation. The code is available at the author's GitHub repository `https://github.com/xunzheng/notears`.

How to compute AUC and SHD on the different baselines

We have mentioned that all comparisons are made using estimated skeletons. The AUC considers a range of precision / recall values estimated with different parameters to determine the presence / absence of edges.

- For DECS and Notears this computation is straightforward as both return weighted adjacency matrices and one obtains a skeleton by choosing different thresholds on the estimated weights to determine presence / absence of edges.

- For LGES the strategy is different as it does not return weighted adjacency matrices. The equivalence class of LGES is computed using the BIC and we obtain a range of precision / recall values by considering a range of penalties on the strength of the BIC regularization, as done by the authors in [58].
- FCI uses independence tests to recover the skeleton and thus requires a threshold for significance, precision / recall values are obtained by varying this threshold.

E.2.2 Further experiments with sparse unobserved confounding

We conduct in this section an empirical investigation on the sensitivity of DECS with respect to the level of denseness on B. Sparse unobserved confounding render the spurious contributions to the adjacency matrix indistinguishable from the true causal signal. We consider B to be drawn as W in the data generating mechanism, i.e. a DAG with a specified number of edges e (fewer edges implying sparser unobserved confounding contribution).

We evaluate all algorithms on the Gaussian model with Erdös–Rényi and $p = 20$ nodes with the difference that B is drawn as W with e edges (recall that W has fixed $e = 20$ edges).

As can be seen in Table E.1, with decreasing number of non-zero entries in B, that is increasing sparsity, the advantage of DECS decreases, though performance remains competitive.

Table E.1: SHD as a function of the number of non-zero entries in B

	20	**50**	**100**	**200**
DECS	62 ± 6.0	53 ± 7.3	50 ± 6.2	46 ± 5.9
NOTEARS	68 ± 3.9	71 ± 2.5	70 ± 2.8	75 ± 2.0
LGES	56 ± 3.3	60 ± 5.1	58 ± 6.0	53 ± 3.4
FCI	67 ± 6.2	85 ± 3.7	85 ± 5.0	95 ± 5.7

E.2.3 Further experiments using adjacency matrix error

In the main body of this paper we tested performance on undirected graphs to allow for comparisons across algorithms with different outputs. Here we consider recovery performance of the original weighted adjacency matrix W used to generate the data. Comparisons are made with Notears which is the only method that returns a weighted adjacency matrix although it does not account for unobserved confounding. This experiment thus served to show that

adjusting for unobserved confounding can significantly improve upon the same algorithm without adjustments.

We follow the same experimental set-up as in the main body of this paper and report results in Figure E.1.

E.2.4 Further reproducibility experiments on skeleton recovery

In the main body of this paper we tested for the reproducibility of causal discovery in different environments shifted by the distribution of unobserved confounders. In this section we consider the exact same set-up but test instead for skeleton recovery to be able to make comparisons with LGES.

Results are given in Figure E.2. The results show that DECS returns a skeleton which is more reproducible across environments. For instance, approximately 20% of estimated edges in the skeleton (across all 10 environments) intersect in all 10 environments for DECS whereas only 7% and 3% do for LGES and Notears respectively.

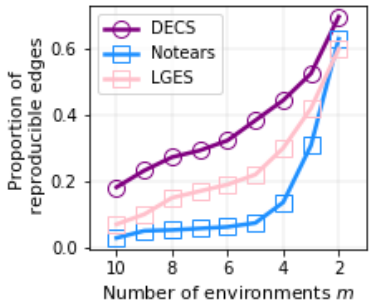
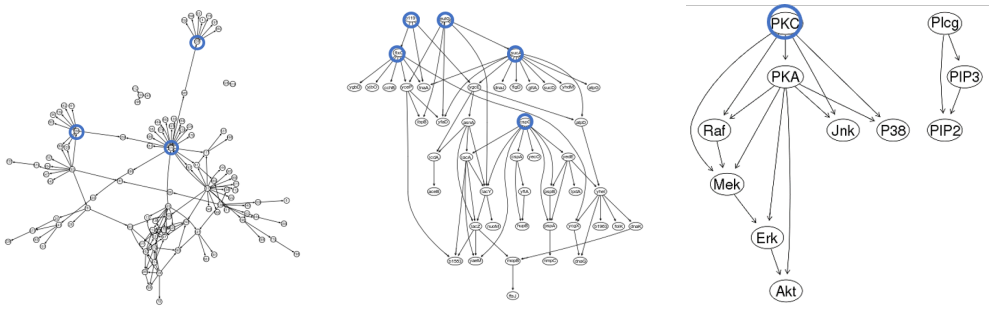


Figure E.2: Reproducibility experiments on skeleton recovery. Higher values for larger number of environments indicate higher levels of reproducibility. DECS is the proposed approach.

E.3 Details on Genetic (semi-synthetic) data

- The **Scale Free (SF)** graph is simulated using the Barabási-Albert model [7], which is based on the preferential attachment process, with nodes being added sequentially. In particular, 1 edge is added each time between the new node and existing nodes. Scale-free graphs are popular since they exhibit topological properties similar to real-world networks such as gene networks, social networks, and the internet. Once the network G is sampled we draw edge weights and data following the Erdös-Rényi data generating process with $n = 100, p = 200, q = 10, \sigma = 0.2$.
- The **E. coli** network describes the expression of protein coding genes of the E. coli microorganism under stress, in an experiment conducted by [152]. The available data of 100 samples of 46 genes was sampled from a Gaussian model, as described in the `bnlearn` R package.
- The **Starch** network simulates gene expression expression interaction resulting from an experiment investigating the impact of the diurnal cycle on the starch metabolism of Arabidopsis thaliana [125]. This gene network and data contains 107 genes, 150 edges and 100 samples and represents an example of a high-dimensional causal discovery problem. It is available in the `bnlearn` R package.
- The **Sachs** dataset consists of $n = 7466$ measurements of expression levels of proteins and phospholipids in human immune system cells for $p = 11$ cell types [148]. It is widely used as a benchmark for causal discovery as it comes with a consensus network that is accepted by the biological community. It is available in the `bnlearn` R package.

We give illustrations of the real networks, together with omitted nodes in Figure E.3. Variables in blue are root nodes omitted from the available data to induce unobserved confounding among children, and thus simulate a scenario of incomplete system of variables as would be expected in real applications.



(a) Starch network (b) E. coli network (c) Sachs network

Figure E.3: Networks and omitted variables considered in the genetic data experiments.