The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Identifying Selection Bias from Observational Data

David Kaltenpoth, Jilles Vreeken

CISPA Helmholtz Center for Information Security, Germany
david.kaltenpoth@cispa.de, vreeken@cispa.de

Abstract

Access to a representative sample from the population is an assumption that underpins all of machine learning. Unfortunately, selection effects can cause observations to instead come from a subpopulation, by which our inferences may be subject to bias. It is therefore essential to know whether or not a sample is affected by selection effects. We study under which conditions we can identify selection bias and give results for both parametric and non-parametric families of distributions. Based on these results, we develop two practical methods to determine whether or not an observed sample comes from a distribution subject to selection bias. Through extensive evaluation on synthetic and real-world data, we verify that our methods beat the state of the art both in detecting as well as characterizing selection bias.

Introduction

In order to draw valid conclusions about the underlying probability distribution, statistical learning theory assumes that we have access to a representative sample from the population. Selection effects, induced by preferential inclusion of some data based on unknown factors causally downstream of the observed variables, violate this assumption and cause what is known as selection bias.

As an example, consider the study by Kovács and Sharkey [2014] on Goodreads book ratings. In a sample of 32 books, they found that the average ratings for individual books went down after winning an award. This is explained by the fact that there are two kinds of readers. Those who read a book before it won an award did so because they were predisposed to like it and therefore more likely to give good ratings in the first place. Meanwhile, those who read books after they won an award were not predisposed to like the book and are likely to be more representative of the population as a whole.

Similar and less benign issues occur in other empirical sciences, such as case-control studies in epidemiology [Glymour and Greenland 2008], studies using hospital-admission data [Berkson 1946, Herbert et al. 2020], genetics [Mefford and Witte 2012], economics [Angrist 1997] and in statistics [Kuroki and Cai 2006].

In machine learning, predictive performance suffers from selection bias when training samples are collected preferentially, as this leads to covariate shift between training and test samples [Bickel, Brückner, and Scheffer 2009]. Methods that correct for this shift [Sugiyama, Krauledat, and Müller 2007, Gretton et al. 2009, Mallick et al. 2022] rely heavily on the availability of independent training and test datasets which can be matched with each other to correct for such distribution shifts. Therefore, they cannot tell us whether selection bias might be affecting our sample when we have access only to a single dataset.

In contrast, in this work, we study conditions under which selection bias is identifiable, given only a single dataset. We provide identifiability results for both parametric and non-parametric distributions when the selection effect is linear.

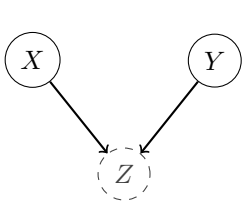
For the former, we provide results for exponential families both when the selection is a deterministic function of observed covariates as well as when additive Gaussian noise may influence the selection of data points. For the latter, we show identifiability under the general assumption that the distribution is subject to certain subsets of invariances, such as rotational symmetries or quantile inversions.

Based on these theoretical results, we propose two practical methods to tell whether data from only a single dataset has been subjected to selection bias, as well as how strong this bias is. That is, we can estimate where the selection boundary lies and how much of the original distribution has been lost to selection bias.

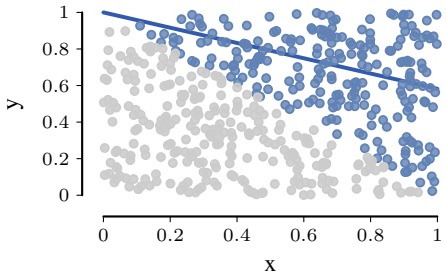
Through an extensive set of experiments, including case studies on penguin and exoplanet data, we show that our methods provide useful and novel insight as well as significantly outperform baselines that measure distribution shifts or model confounding factors.

Our contributions are as follows:

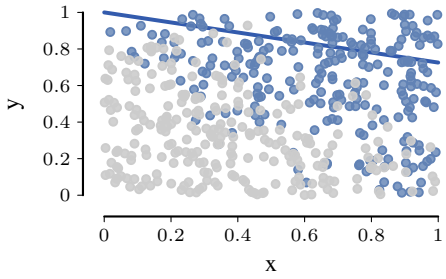
1. **Identifiability of selection bias:** We provide theoretical results on the identifiability of selection bias in both parametric and non-parametric model families.
2. **Learning orthogonal symmetries:** We provide a method that learns special orthogonal symmetries of the data-generating distribution in an unsupervised manner by using the Cayley transform.
3. **Detecting selection bias from a single dataset:** We provide two algorithms to recover the selection boundary given a single dataset and estimate the strength and effect of the selection bias.



(a) Selection bias in a graphical model



(b) Selection on $X + Y > 1$.



(c) Selection on $X + Y + \epsilon > 1$.

Figure 1: Selection bias. Left: A graphical representation of selection bias as the act of conditioning on a common child Z of multiple observed variables. Middle: The effect of selection on independently uniformly sampled points. Blue points are observed, gray points are excluded due to selection. While X and Y are originally independent, they are negatively correlated in the observed sample. Right: Similar to (b) but the selection is noisy. Included and excluded points are no longer nicely separated, making it more difficult to notice the effect of selection.

Our paper is structured as usual. We begin by formally defining our problem in the next section. Then, in Section , we give our identifiability results. Next, in Section , we describe our proposed methods. Last, in Section , we empirically evaluate our methods before wrapping up in Section . We make all code and data available online.![1]

Preliminaries

In this section, we begin by providing common notation. We then formally define selection bias and state assumptions on the precise kind of selection bias we consider in this paper.

Notation

We denote random variables by capitals, e.g., X, Y, Z, where we write $X = (X_1, \ldots, X_n), Y$ for observed, and Z for unobserved variables. We denote samples from these by small letters, e.g., x, y, z. For ease of notation when writing affine-linear expressions $a^\top X + a_0$, we add a variable $X_0 = 1$ so that the expression becomes simply $a^\top X$. We denote sets of vectors a by $\mathcal{A}$ and index sets by I and write $[n] = \{1, \ldots, n\}$.

We write probability distributions as P, Q, with densities p, q whenever they exist. We generally use Q to refer to distributions $P(\cdot \mid Z \in A)$ for a set A, e.g. $P(\cdot \mid a^\top X > 0)$.

Vector norms are denoted by $\|v\|$ and transposes by $v^\top$.

Selection Bias

Given observed variables X, we define selection bias as the act of conditioning on an unobserved variable Z caused by X. This causes a distribution shift from the population distribution $P(X)$ to a distribution $P(X \mid Z)$, resulting in potentially false inferences upon $P(X)$. We consider the most general case, where $Z = f(X, \epsilon)$ can be a function of any or all variables in X as well as some independent noise term ϵ.

As an example, consider the setup shown in Fig. 1. We let $P(X, Y) = U[0, 1]^2$, making them independent. We then select on $X + Y > 1$, leaving only the data in the top right. We see clearly that there is a negative linear relationship in the distribution $Q(X, Y) = P(X, Y \mid X + Y > 1)$.

We call the type of selection above *noiseless* because it depends solely on X, Y but no external source of noise. Selection can also be *noisy*, e.g., $X + Y + \epsilon > 1$, where $\epsilon \sim N(0, 0.05)$ is a small amount of noise. We show this in Fig. 1 on the right. Clearly, noise makes it more difficult to determine whether selection is occurring and, if so, where.

For example, some studies early in the COVID-19 pandemic found that smoking appeared to be protective against lung cancer [Herbert et al. 2020]. This turned out to be, in part, due to the use of hospital admission data, and created a classical case of Berkson's Paradox [Berkson 1946]: people generally go to the hospital for *some* reason, and when one possible cause is ruled out, the others become more likely.

In this paper we assume linear selection $Z = f(X, \epsilon) = a^\top X + \epsilon$ so that our observed sample x comes from the distribution $Q_a = P(\cdot \mid a^\top X + \epsilon > 0)$ where $a \in \mathcal{A}$ is unknown. We further assume that each vector $a \in \mathcal{A}$ is normalized, e.g., $a_1 = 1$. Our goal then is to recover a and P from a sample $x \sim Q_a$. Next, we describe the theoretical underpinnings of two different approaches to doing so.

Theory

In this section, we show that linear noiseless selection effects are always identifiable for exponential families, and noisy selection with Gaussian noise is identifiable for the normal family. Further, for non-parametric families, we show that linear noiseless selection is identifiable under assumptions on the invariances to which P is subject.

We include further details on the theory as well as proofs for all results in the appendix.

Selection Bias in Parametric Models

We call a set $\mathcal{M}$ of probability distributions P *parametric* if each distribution P can be parameterized (uniquely) by a finite-dimensional vector of parameters $\theta \in \Theta$.

[1] https://eda.mmci.uni-saarland.de/prj/sprite/

An example of this are *exponential families*, where each P_θ has density p_θ defined as [Bishop and Nasrabadi 2006]

$$p_\theta(x) = h(x)\exp(\eta(\theta)^\top T(x) - A(\theta)),$$

where $\eta(\theta)$ are called the natural parameters, $T(x)$ the sufficient statistic and $A(\theta)$ the log-partition function. Further, note that by definition all P_θ share the same support $\text{supp}(P_\theta) \coloneqq \{x : p_\theta(x) > 0\}$. Uniform distributions over different sets, therefore, do not form an exponential family.

With selection bias in play we form the model class $\mathcal{M}_s = \{Q_{\theta,a} : P_\theta \in \mathcal{M}, a \in \mathcal{A}\}$ where $Q_{\theta,a} \coloneqq P_\theta(\cdot \mid a^\top X + \epsilon > 0)$. Note that $\mathcal{M}_s$ no longer forms an exponential family since its members do not share the same support.

Next, we move on to identifiability results for the parameters of the distributions $Q_{\theta,a} \in \mathcal{M}_s$.

Identifiability for Exponential Families

We begin by showing that noiseless selection is always identifiable for exponential families as long as the set $\mathcal{A}$ of selection coefficients a does not contain degenerate values that result in selecting either all or no samples.

Theorem 1 (Identifiability for noiseless selection in exponential families)**.** *Let $\mathcal{M}$ be an exponential family with parameter space Θ and $\mathcal{A}$ the set of a such that*

$$0 < P_\theta(a^\top X > 0) < 1$$

for all θ. Then the parameters (θ, a) of $Q_{\theta,a}$ are identifiable.

The assumption that $0 < P_\theta(a^\top X > 0) < 1$ is natural. First, if $P_\theta(a^\top X > 0) = 0$ then the distribution $Q_{\theta,a}$ is not well-defined. Conversely, if $P_\theta(a^\top X > 0) = 1$, then no actual selection occurs, and the same would be true for any $a_0' > a_0$, making the parameters non-identifiable.

Next, we prove identifiability in the case of noisy selection in the Gaussian exponential family.

Theorem 2 (Identifiability for noisy selection in the Gaussian family)**.** *Let $\mathcal{M}$ be the Gaussian exponential family with parameter space $\Theta = \{(\mu, \Sigma)\}$ and let $\epsilon \sim N(0, 1)$. Further, let*

$$Q_{\mu,\Sigma,a,\zeta}(X) = P_{\mu,\Sigma}(X \mid a^\top X + \zeta\epsilon > 0).$$

Then the parameters $(\mu, \Sigma) \in \Theta, a \in \mathbb{R}^{m+1}, \zeta > 0$ are jointly identifiable, where a is normalized as stated above.

The assumption that $0 < P_\theta(a^\top X > 0) < 1$ is not necessary here because $\text{supp}(P_\theta) = \mathbb{R}^m$.

Next, we consider the case where our distributions are no longer necessarily of a known parametric form but instead satisfy another regularity condition by way of invariances.

Beyond Parametrics: Invariance

Assume that we have data either from a normal distribution $N(\mu, \sigma^2)$ or from a t-distribution $t_\nu(\mu, \Sigma)$ with ν degrees of freedom. Then, while we do not know the model class from which our data comes, we nevertheless know one crucial fact about the underlying distribution: it is the same after reflection across its mean, i.e., X and $-X + 2\mu$ have the same distribution. We call this an invariance of the distribution, which we define formally next.

Definition 1 (Invariance)**.** *Let P be a probability distribution and j be a measurable bijective function. We say that P has invariance j if $P(j(A)) = P(A)$ for all measurable A.*

Note that the function j can be arbitrarily complex and fine-tuned to P. For example, the 1-dimensional exponential distribution $\text{Exp}(\lambda)$ has the invariance $t \mapsto -\log(1 - e^{-\lambda t})/\lambda$, which is simply the mapping of its q-quantile to its $1-q$-quantile. In fact, every univariate distribution P with connected support has such a *quantile inversion* mapping of q to $1-q$-quantiles. If $F(t) = P(X \leq t)$ then this is the mapping $t \mapsto F^{-1}(1 - F(t))$. Furthermore, by Sklar's theorem [Sklar 1959, Jaworski et al. 2010], every multi-variate distribution $P(X)$ with connected support has an invariance group generated by its marginal quantile-inversion maps.

If P has a density p then for any x_0, x_1 it has trivial invariances $j(x_i) = x_{1-i}$ and $j(x) = x$ everywhere else. Such invariances are said to be equal to the identity P-almost everywhere. To preclude such degenerate cases, we consider what we call *strongly distinguishable invariances*.

Definition 2 (Strongly distinguishable invariances)**.** *A set J of invariances is called strongly distinguishable for P if for all $j, j' \in J$ we have $P(j(x) = j'(x)) > 0$ iff $j = j'$.*

Consider, for example, the normally distributed $X \sim N(0, \sigma^2 I)$. It is invariant under any orthogonal matrix U, since $UX \sim N(0, \sigma^2 UU^\top) = N(0, \sigma^2 I)$. Further, for any two $U \neq U'$, the set $K = \ker(U - U') = \{x : Ux = U'x\}$ is a linear subspace of $\mathbb{R}^m$ with $\dim(K) < m$ so that $P(UX = U'X) = 0$. The set of orthogonal matrices U is therefore strongly distinguishable for any $N(0, \sigma^2 I)$.

We can use these invariances to detect selection bias. If P is invariant under j, then selection bias will likely break such an invariance. For example, if we disregard all samples to the right of the 80% quantile of the exponential distribution, then the new distribution Q will no longer be invariant with respect to the map above. The invariance j still applies to large parts of Q, however, indicating that we can nevertheless obtain useful information about P. We formalize this intuition in the following theorem.

Theorem 3 (Identifiability of Selection under Invariance)**.** *Let $\mathcal{M}$ be a set of probability distributions and J be strongly distinguishable for each $P \in \mathcal{M}$. Assume that for all $P \in \mathcal{M}$ there is $j \in J$ such that $P(X) = P(j(X))$. Let $P \in \mathcal{M}$ and $\mathcal{A}$ be the set of a such that*

$$0 < P(a^\top X \leq 0) < 1/2.$$

Then a is identifiable. Further, if all distributions $P_1, P_2 \in \mathcal{M}$ satisfy $P_1(\cdot \mid a^\top X > 0) = P_2(\cdot \mid a^\top X > 0)$ iff $P_1 = P_2$ then P is identifiable too.

The last assumption is true for many classes of distributions. In particular, it holds for all exponential families, unions of multiple exponential families, arbitrary finite mixtures of exponential families, and stationary Gaussian processes [Bishop and Nasrabadi 2006].

Methods

This section develops two complementary approaches to discovering selection bias in observational data. The first,

which we will refer to as EXP, directly fits an exponential family with selection bias to the data. The second, referred to as INV, finds an approximate invariance of the true distribution and derives a likely selection boundary from it.

Finding Selectors for Exponential Families

When we know that the data x comes from an exponential family (potentially) subject to selection bias, the simplest approach is to estimate the parameters θ, a, ζ of $Q_{\theta,a,\zeta}$. The log-likelihood is given by

$$\begin{aligned} l_{\theta,a,\zeta}(x) &:= \log q_{\theta,a,\zeta}(x) \\ &= \left(\sum_{i=1}^{n} \log p_\theta(x_i) + \log p_\zeta(a^\top x_i + \epsilon > 0)\right) \\ &\quad - n \log P_{\theta,\zeta}(a^\top x + \epsilon > 0) \ . \end{aligned}$$

To obtain a good set of parameters, we suggest updating θ, a, and ζ alternately as follows. Starting from a random initialization θ_0, a_0, ζ_0, we can update our parameters as follows

$$\begin{aligned} \theta_{t+1} &\leftarrow \theta_t + \lambda_\theta \partial_\theta l_{\theta_t,a_t,\zeta_t}(x), \\ a_{t+1} &\leftarrow a_t + \lambda_a \partial_a l_{\theta_{t+1},a_t,\zeta_t}(x), \\ \zeta_{t+1} &\leftarrow \zeta_t + \lambda_\zeta \partial_\zeta l_{\theta_{t+1},a_{t+1},\zeta_t}(x), \end{aligned}$$

with step sizes $\lambda_\theta, \lambda_a, \lambda_\zeta$. We give them the full gradient $\nabla l_{\theta,a,\zeta}$ in the appendix. Note that, as with exponential families in general, we cannot provide convergence guarantees for our approach. We will see, however, that we empirically obtain good estimates.

To measure the gain of modeling the selection boundary, we further compute the confidence of our method as

$$0 < C = \frac{l_{\theta,0,0}(x) - l_{\theta,a,\zeta}(x)}{l_{\theta,0,0}(x)} < 1$$

It measures the relative improvement from including selection into the modeling process over the model where lack of selection is assumed. The confidence is low when no gains can be made from modeling selection, and it is high when large gains can be made. We will see in Section that it correlates well with the empirical performance of our approach.

Finding Invariances and Selection Boundaries

Next, we move away from strict model assumptions and instead develop an approach based on invariances of the underlying distribution as outlined in Sec . For the sake of feasibility, we restrict ourselves to the simple but still expressive class of orthogonal matrices—but note that recently some progress on discovering more general symmetries has been made [Desai, Nachman, and Thaler 2021].

To motivate our approach, recall that if P is invariant under U, the sample Ux should look indistinguishable from the sample x. Hence, given sample x from the distribution $Q_a = P(\cdot \mid a^\top X > 0)$, we would like to maximize some similarity measure of the datasets Ux and x. Since selection is at play, however, even the true invariance U cannot apply to all samples x_i.

To address this issue, we will first have to introduce our approach based on the kernel mean embedding μ_P of P [Muandet et al. 2017]. One can show under relatively general conditions that $\mu_P = \mu_Q$, iff $P = Q$ [Muandet et al. 2017]. In particular, $\mu_P = \mu_{P \circ U}$ iff U is an invariance of P.

In the absence of selection bias, our goal would be to find the matrix U that minimizes $|\mu_P - \mu_{P\circ U}|^2$. The empirical estimate for a sample x is [Muandet et al. 2017]

$$\begin{aligned} \frac{1}{N}\sum_{i,j} k(x_i, x_j) + \frac{1}{N}\sum_{i,j} k(Ux_i, Ux_j) \\ - \frac{1}{N^2}\sum_{i,j} k(Ux_i, x_j) \geq 0. \end{aligned}$$

In practice, it is useful to use an isotropic kernel, i.e., $k(Ux, Uy) = k(|Ux - Uy|) = k(|x - y|)$. Fortunately, the commonly used squared exponential kernel $k(x, y) = \exp(-\lambda |x - y|^2)$ satisfies this property. Then the first two terms above are independent of U so that we maximize

$$L(U; x) = \frac{1}{N^2}\sum k(Ux_i, x_j).$$

However, due to the effects of selection bias, not all samples available to us are "good" samples. To deal with this, we use the following approach to determine which samples are good. To begin with, if U^* is an invariance of P, how does the selection mechanism $a^\top X > 0$ affect this?

Clearly, the points for which $a^\top U^* x_i > 0$ are unaffected in the above score. Meanwhile, those points for which $a^\top U^* x_i \leq 0$ are far away from every point x_j in the available sample, incurring a large penalty. Hence, the score L would be improved if we recomputed the average without the terms $k(U^* x_i, x_j)$.

As such, we propose the following approach to determining which samples are "good" samples. First, we optimize $L(U; x)$ with respect to U (described below). Then, for each point $x_k \in x$ we check if for $I = \{1, \cdots, k-1, k+1, \cdots, n\}$ we have

$$L(U; x, I) = \frac{1}{N\,|I|}\sum_{i\in I, j\in[n]} k(Ux_i, x_j) \ll L(U; x).$$

In other words, we check if the set of points $\{Ux_i\}_{i\in I}$ is significantly more similar to the sample x than the sample $\{Ux_i\}_{i\in[n]}$. We then remove all "bad" points and set $I = [n] \setminus \{k_1, \cdots, k_l\}$.

We then rerun the optimization of U starting at its previous optimum, now using the score $L(U; x, I)$ instead. After each such optimization step over U, we evaluate $L(U; x, [n])$ and recompute the set I by removing indices from $[n]$ rather than updating I directly. This is necessary as some previously considered "bad" might only have appeared as such due to U being poorly optimized. We repeat this process until the pair (U, I) converges.

To optimize the score $L(U; x, I)$ over special orthogonal matrices, we use the Cayley transform [Wen and Yin 2013]

$$U = (I - A)(I + A)^{-1}$$

where A is a skew-symmetric matrix, $A^\top = -A$. This turns the constrained optimization problem of $L(U)$ with respect

to orthogonal matrices into an unconstrained optimization problem $L(A)$ with respect to skew-symmetric matrices.

Note that while the matrices U parametrized in this way all lie in $\mathrm{SO}(n)$, i.e., $\det(U) = 1$, this is not a concern for us. If P is invariant to U, then it is also invariant to $U^2 \in \mathrm{SO}(n)$. Since we use the matrix U only to determine the effects of selection bias, this purpose is equally well-served by working only with matrices in $\mathrm{SO}(n)$.

This makes the optimization dramatically more efficient and stable than other approaches, such as coordinate descent using Givens rotations [Shalit and Chechik 2014].

Having obtained an orthogonal matrix U and the index set I, we can use these to estimate the selection boundary. Let $I^c = [n] \setminus I$. Then the points $x_k, k \in I^c$ are such that Ux_k is far from observed samples x_i and is likely to lie in the region $a^\top Ux_k < 0$, i.e., the other side of the selection boundary. We therefore use a linear classifier such as an SVM to separate the two sets of points $\{x_i\}_{i\in[n]}$ and $\{Ux_k\}_{k\in I^c}$. We will see in the experiments that this simple approach already produces good results.

We compute the confidence of this method by

$$C = \frac{L(\widehat{U}; [n]) - L(\widehat{U}; \widehat{J})}{L(\widehat{U}; [n])},$$

again measuring the improvement from modeling selection.

Related Work

While most machine learning and statistical research assumes access to a representative sample from the population, selection bias can have detrimental effects on statistical inferences, especially regarding public health advice [Berkson 1946, Herbert et al. 2020]. Work with samples subject to selection bias can also reinforce stereotypes, causing issues with regard to the fairness of algorithmic decision-making [Caton and Haas 2020].

Most work done on the topic of selection bias focuses on conditions under which selection bias can be controlled for [Bareinboim and Pearl 2012, Bareinboim, Tian, and Pearl 2014, Bareinboim and Pearl 2016, Forré and Mooij 2020, Versteeg, Zhang, and Mooij 2022], or identifiability of causal directions in spite of a special case of selection bias [Zhang et al. 2016]. In contrast, we are concerned with conditions under which it is possible to determine whether selection bias is a likely concern for a given dataset.

Related approaches are those dealing with covariate shift [Gretton et al. 2009, Sugiyama, Krauledat, and Müller 2007]. They require access to multiple datasets, however, making them unusable when only one dataset subject to selection bias is available. Similarly related to selection bias is confounding in causal inference and discovery [Wang and Blei 2019, Kaltenpoth and Vreeken 2019, Bhattacharya et al. 2021]. However, the approaches used here do not transfer to selection bias despite their apparent similarity.

The study of symmetries in probability distributions garnered much attention at the start of the century [Fang, Kotz, and Ng 1990, Chikuse 2003, Kallenberg 2005].

More recent theoretical work has focused chiefly on providing theoretical frameworks to explain the benefits of symmetries for predictive tasks [Lyle et al. 2020, Fortuin 2022, Chen, Dobriban, and Lee 2020, Dao et al. 2019]. A different line of research has focused on learning models invariant to a given symmetry group T. van der Wilk et al. [2018] developed invariant Gaussian processes f by averaging a Gaussian process g over the orbit of T. Further work also extended this line of work to neural networks [van der Ouderaa and van der Wilk 2022]. Note that these approaches assume that T is known beforehand. Benton et al. [2020] relax this assumption by parametrizing the set of transformations. Other recent work has focused on leveraging the benefits of symmetries, especially in image recognition systems [Ravanbakhsh, Schneider, and Poczos 2017, Worrall et al. 2017, Immer et al. 2022]. However, these approaches focus on exploiting symmetries in the data-generating process to achieve a specific supervised task.

One notable exception to this is SymmetryGAN [Desai, Nachman, and Thaler 2021], which uses a generative adversarial network (GAN) approach to learning a linear volume-preserving transformation of the data, which makes it look indistinguishable from the original dataset.

Experiments

In this section we perform a comprehensive experimental analysis of our proposed methods.

We begin by explaining the data generating process and then show that our methods' performance can be predicted from observable quantities. We further evaluate our methods on two real world datasets and show that they provide relevant and novel insight into the data.

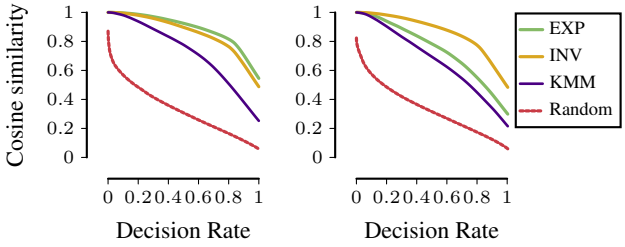
To verify that our methods work, we compare them with two different algorithms. First, we use the kernel mean matching (KMM) [Gretton et al. 2009] algorithm, designed to tell whether there is a distribution shift between two different datasets. Second, we use DCD [Bhattacharya et al. 2021], a recent approach to causal discovery in the presence of confounding. It explicitly models non-causal edges, making it a suitable competitor.

We implement our methods in Python using Tensorflow [Abadi et al. 2016] and use the publicly available implementations of KMM and DCD. All experiments finished within a few hours on a commodity laptop. We make all code and data available in the supplementary material.

Data Generation

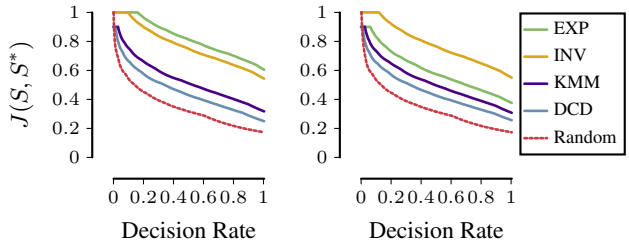
To generate suitable synthetic data, we use the following approach. We start by generating a random directed Erdős-Rényi (ER) network G with probability of an edge being added being p. To do so, we sample a random topological order τ over the nodes $\{1, \ldots, m\}$ and for $i < j$ add an edge $(\tau(i), \tau(j))$ with probability p. We define the distribution over $X_1, \cdots, X_m$ via the structural model $X_i = f_i(\mathrm{pa}_i, \epsilon_i)$ for appropriate functions f_i and noise variables ϵ_i, where pa_i are the parents of X_i in G.

For the multivariate Gaussian distribution, this is $X_i = \beta_i^\top \mathrm{pa}_i + \epsilon_i$ where $\beta_i^\top \sim N(0, \sigma_\beta^2)$ and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. For data generation from the t-distribution we use the approach proposed by Finegold and Drton [2014].



(a) Normally distributed data (b) t-distributed data

Figure 2: [Higher is better.] Decision rate plots of the cosine similarity between discovered and true selection boundary. Left: Experiments on Gaussian distributed data. EXP outperforms INV slightly, and both outperform KMM significantly. Right: Experiments on t-distributed data. INV outperforms both EXP and KMM significantly. In both cases, all methods significantly outperform random guessing.



(a) Normally distributed data (b) t-distributed data

Figure 3: [Higher is better.] Decision rate plots of the Jaccard similarity between recovered and true set of variables affected by selection bias. Left: Experiments on Gaussian distributed data. EXP outperforms INV slightly, and both outperform KMM and DCD significantly. Right: Experiments on t-distributed data. INV outperforms EXP, which in turn outperforms KMM and DCD significantly. In both cases, all methods significantly outperform random guessing.

We generate samples $x = (x_1, \ldots, x_m)$ from $P(X)$ and then pick a random sink node Z from G and remove all samples for which $Z + a_0 < 0$ where $a_0 \sim N(0, \sigma^2(f_Z))$. For the Gaussian distribution, we pick $\sigma^2(f_Z) = \beta_{i_1}^2$ where i_1 is the first parent of Z in the topological ordering. Note that this is the setting used in Thm. 2 if $P(X)$ is Gaussian.

For each instantiation of the parameters, we generate data points until a total of 1000 points are included in the observed data. We further run each experiment 1000 times to obtain reliable results.

Recovering the Selection Boundary

We start our evaluation by checking how well each method predicts the correct selection boundary in a dataset that is subject to selection bias. To this end, we generate data from three-dimensional Gaussian and t-distributions $N(\mu, \Sigma)$ and $t_\nu(\mu, \Sigma)$ subject to selection bias as described above, where we set $p = 1$ for our ER network. We consider the impact of larger dimensions in the appendix.

We then compute the cosine similarity $0 \leq \frac{\langle a, a^* \rangle}{|a||a^*|} \leq 1$ between the true selection boundary a^* and the estimated a. A result closer to 1 corresponds to better performance.

We compare only with KMM here as DCD is not capable of estimating a^*. For KMM, we summarize here the modifications made to make it applicable to our setting. Full motivation and details are included in the appendix.

Since KMM requires two datasets, besides the one subject to the true selection boundary a^*, we also give it access to a second dataset subject to the selection boundary a' which is a slightly rotated version of a^*. Thus, the original data x and the secondary dataset x' share similar distributions which are nevertheless different and are therefore amenable to analysis using KMM.

We show the results in a decision-rate plot in Fig. 2. On the x-axis is the decision rate, i.e. the fraction or number of datasets evaluated so far, ordered from most to least confident for each method. On the y-axis is the cosine similarity. We see clearly that for all three methods, the confidence strongly correlates with their performance on both datasets.

On the left, we see that for Gaussian generated data, EXP performs slightly better than INV, although not significantly. Both methods significantly outperform KMM. On the right, for t-distributed data, INV significantly outperforms EXP which in turn significantly outperforms KMM. Lastly, all methods significantly outperform random guessing of the selection boundary on both datasets at all levels.

Recovering Variables Affected by Selection

Next we consider the task of discovering which variables are affected by selection bias. We generate data from a ten-dimensional joint distribution with $p = 0.3$ for the ER Graph as described above. Then the parents of the variable Z we condition on are the variables we would like to recover.

For evaluation, we compute the Jaccard similarity between the true set $S^* = \text{pa}_G(Z)$ and our recovered S,

$$J(S, S^*) = \frac{|S \cap S^*|}{|S \cup S^*|},$$

where higher values tell us that S is more similar to S^*. We include further analysis of the precision-recall curve for this experiment in the appendix.

We compare our methods with both KMM and DCD in this setting. For our methods and KMM, we use the discovered selection boundaries a and consider those variables which have a_i significantly different from zero to be the variables subject to selection bias. For DCD, we run the method to obtain pairs of variables whose correlations are estimated to be strictly non-causal. We then estimate the set of variables affected by selection to be the set of all variables included in at least one such pair.

We show the resulting decision rate plots in Fig. 3. As in the previous section, for Gaussian generated data, EXP outperforms INV slightly. Further, both of our methods outperform both KMM and DCD significantly. For t-distributed data, INV again outperforms both EXP significantly, which

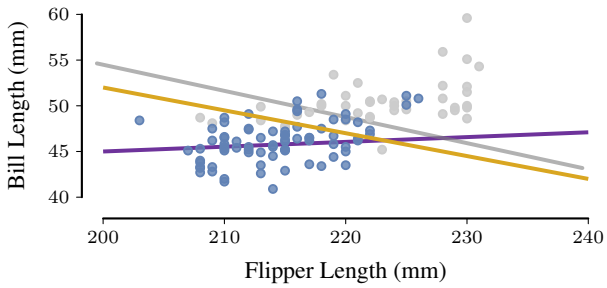


Figure 4: Result on Palmer penguin dataset. Observed points are displayed in blue, unobserved points in gray. The linear separator found by LDA is displayed in gray. We see that the selection boundary found by INV (gold) is much closer to the best possible than the one found by KMM (purple).

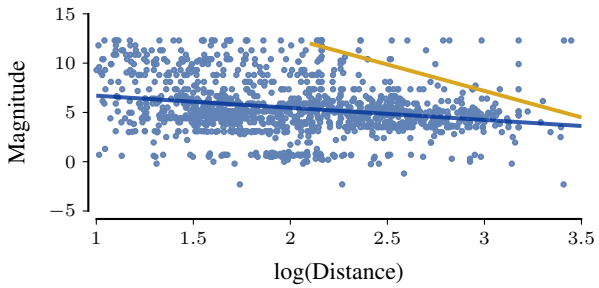


Figure 5: Result on exoplanet dataset. The regression line (dark blue) is negative and highly significant, $p < 10^{-20}$. The selection boundary estimated by INV (gold) captures the intuitive lack of points in the top right corner, consistent with our speculation of selection due to technological limitations.

in turn significantly outperforms both KMM and DCD. Further, all methods significantly outperform random guessing of the set of variables affected by selection.

Real Data

To see if our methods can provide novel insight, we evaluate them on two real-world datasets.

Palmer Penguins We begin by evaluating our methods on the Palmer Penguins dataset [Gorman, Williams, and Fraser 2014]. They were collected at Palmer Station, Antarctica, and contain samples from three different species of penguins. Among the measured variables are bill depth, bill length, flipper length, and weight.

To test whether our approach is capable of finding interesting results on this dataset, we preprocess our data as follows. We split the data by penguin species and then for each of them we select the 80% of penguins with the lowest weight from that species. Since we expect bill depth, bill length and flipper length to all positively affect weight, this should introduce selection bias in these features.

We show an example of our results in Fig. 4, where gray points have been removed in the preprocessing step. We see that the selection boundaries estimated by our approach is reasonable, while the one provided by KMM is not. In fact, when compared to a Linear Discriminant Analysis (LDA) [Bishop and Nasrabadi 2006] fit on *all* data with known labels of which data points our methods have access to, the selection boundaries discovered by our methods are almost identical. We show similar results for other pairs of variables and penguin species in the appendix.

Exoplanet Discovery Next, we consider data from the Open Exoplanet Catalogue using the ExoData library [Rein 2012, Varley 2016]. It contains data about exoplanets and their stars, including variables such as their distance d from the earth and their absolute magnitude – a measure of their brightness, with dim stars having high magnitude.

It is generally believed that the universe is uniform at large scales [Liddle 2015]. In particular, the distance of stars from us should be independent of their absolute magnitude. However, due to technological constraints we should expect that the further away a star, the brighter it has to be for us to be able to detect exoplanets in its system. As such, we expect selection effects to be at work in this dataset, making it a good case study for our methods.

We show the data in Fig. 5. One thing that stands out from the very first glance is that the top right corner of our dataset (dim points which are very far away) is only sparsely populated. Indeed, the linear correlation between $\log(d)$ and magnitude is negative and significant at the 10^{-21} level. Applying INV we obtain the selection boundary seen in Fig. 5. We see that the selection boundary found by INV is consistent with our speculations of selection effects based on technological limitations.

Discussion and Conclusion

We introduced two different approaches to identifying selection bias from purely observational data. The first is based on the uniqueness of members of exponential families over any set of non-zero probability. The second is based on invariances of the true generating distribution before selection. Theoretically, we prove both approaches to identify selection bias under general conditions. Empirically, we show that both approaches produce good results both for discovering the true selection boundary as well as recovering the set of variables affected by selection bias.

Two important future work directions, and current limitations, are using broader classes of invariances and richer selection models. These challenges can be roughly broken down into three parts: first, parametrize the underlying invariance and learn a model respecting specific invariances, e.g., using normalizing flows [Kobyzev, Prince, and Brubaker 2020] or GANs [Desai, Nachman, and Thaler 2021]. Second, find those subsets of data points on which we obtain the most coherent invariances as well as those where they are least consistent with the available data. Third, determine the selection mechanism which would produce data similar to our observations given the generative process resulting from the previous step. To go beyond linear selection mechanisms, the use of both kernel-based, as well as methods based on latent embeddings such as VAEs [Kingma and Welling 2019] are promising next steps of exploration.

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: A system for large-scale machine learning. In *OSDI16*, 265–283.

Angrist, J. D. 1997. Conditional independence in sample selection models. *Economics Letters*, 54(2): 103–112.

Bareinboim, E.; and Pearl, J. 2012. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, 100–108. PMLR.

Bareinboim, E.; and Pearl, J. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27): 7345–7352.

Bareinboim, E.; Tian, J.; and Pearl, J. 2014. Recovering from selection bias in causal and statistical inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Benton, G.; Finzi, M.; Izmailov, P.; and Wilson, A. G. 2020. Learning invariances in neural networks from training data. *Advances in neural information processing systems*, 33: 17605–17616.

Berkson, J. 1946. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3): 47–53.

Bhattacharya, R.; Nagarajan, T.; Malinsky, D.; and Shpitser, I. 2021. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, 2314–2322. PMLR.

Bickel, S.; Brückner, M.; and Scheffer, T. 2009. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9).

Bishop, C. M.; and Nasrabadi, N. M. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Caton, S.; and Haas, C. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.

Chen, S.; Dobriban, E.; and Lee, J. H. 2020. A group-theoretic framework for data augmentation. *The Journal of Machine Learning Research*, 21(1): 9885–9955.

Chikuse, Y. 2003. *Statistics on special manifolds*, volume 174. Springer Science & Business Media.

Dao, T.; Gu, A.; Ratner, A.; Smith, V.; De Sa, C.; and Ré, C. 2019. A kernel theory of modern data augmentation. In *International Conference on Machine Learning*, 1528–1537. PMLR.

Desai, K.; Nachman, B.; and Thaler, J. 2021. SymmetryGAN: Symmetry Discovery with Deep Learning. *arXiv preprint arXiv:2112.05722*.

Fang, K.-T.; Kotz, S.; and Ng, K. W. 1990. *Symmetric multivariate and related distributions*. Chapman and Hall/CRC.

Finegold, M. A.; and Drton, M. 2014. Robust graphical modeling with t-distributions. *arXiv preprint arXiv:1408.2033*.

Forré, P.; and Mooij, J. M. 2020. Causal calculus in the presence of cycles, latent confounders and selection bias. In *Uncertainty in Artificial Intelligence*, 71–80. PMLR.

Fortuin, V. 2022. Priors in bayesian deep learning: A review. *International Statistical Review*.

Glymour, M. M.; and Greenland, S. 2008. Causal diagrams. *Modern epidemiology*, 3: 183–209.

Gorman, K. B.; Williams, T. D.; and Fraser, W. R. 2014. Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis). *PloS one*, 9(3): e90081.

Gretton, A.; Smola, A.; Huang, J.; Schmittfull, M.; Borgwardt, K.; and Schölkopf, B. 2009. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4): 5.

Herbert, A.; Griffith, G.; Hemani, G.; and Zuccolo, L. 2020. The spectre of Berkson's paradox: Collider bias in Covid-19 research. *Significance*, 17(4): 6–7.

Immer, A.; van der Ouderaa, T. F.; Fortuin, V.; Rätsch, G.; and van der Wilk, M. 2022. Invariance Learning in Deep Neural Networks with Differentiable Laplace Approximations. *arXiv preprint arXiv:2202.10638*.

Jaworski, P.; Durante, F.; Hardle, W. K.; and Rychlik, T. 2010. *Copula theory and its applications*, volume 198. Springer.

Kallenberg, O. 2005. *Probabilistic symmetries and invariance principles*, volume 9. Springer.

Kaltenpoth, D.; and Vreeken, J. 2019. We are not your real parents: Telling causal from confounded using mdl. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, 199–207. SIAM.

Kingma, D. P.; and Welling, M. 2019. An introduction to variational autoencoders. *CoRR*, abs/1906.02691.

Kobyzev, I.; Prince, S. J.; and Brubaker, M. A. 2020. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11): 3964–3979.

Kovács, B.; and Sharkey, A. J. 2014. The paradox of publicity: How awards can negatively affect the evaluation of quality. *Administrative science quarterly*, 59(1): 1–33.

Kuroki, M.; and Cai, Z. 2006. On recovering a population covariance matrix in the presence of selection bias. *Biometrika*, 93(3): 601–611.

Liddle, A. 2015. *An introduction to modern cosmology*. John Wiley & Sons.

Lyle, C.; van der Wilk, M.; Kwiatkowska, M.; Gal, Y.; and Bloem-Reddy, B. 2020. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*.

Mallick, A.; Hsieh, K.; Arzani, B.; and Joshi, G. 2022. Matchmaker: Data Drift Mitigation in Machine Learning for Large-Scale Systems. *Proceedings of Machine Learning and Systems*, 4: 77–94.

Mefford, J.; and Witte, J. S. 2012. The covariate's dilemma. *PLoS genetics*, 8(11): e1003096.

Muandet, K.; Fukumizu, K.; Sriperumbudur, B.; Schölkopf, B.; et al. 2017. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2): 1–141.

Ravanbakhsh, S.; Schneider, J.; and Poczos, B. 2017. Equivariance through parameter-sharing. In *International Conference on Machine Learning*, 2892–2901. PMLR.

Rein, H. 2012. A proposal for community driven and decentralized astronomical databases and the Open Exoplanet Catalogue. *arXiv preprint arXiv:1211.7121*.

Shalit, U.; and Chechik, G. 2014. Coordinate-descent for learning orthogonal matrices through Givens rotations. In *International Conference on Machine Learning*, 548–556. PMLR.

Sklar, M. 1959. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8: 229–231.

Sugiyama, M.; Krauledat, M.; and Müller, K.-R. 2007. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5).

van der Ouderaa, T. F.; and van der Wilk, M. 2022. Learning invariant weights in neural networks. *arXiv preprint arXiv:2202.12439*.

van der Wilk, M.; Bauer, M.; John, S.; and Hensman, J. 2018. Learning invariances using the marginal likelihood. *Advances in Neural Information Processing Systems*, 31.

Varley, R. 2016. ExoData: A Python package to handle large exoplanet catalogue data. *Computer Physics Communications*, 207: 298–309.

Versteeg, P.; Zhang, C.; and Mooij, J. M. 2022. Local Constraint-Based Causal Discovery under Selection Bias. *arXiv preprint arXiv:2203.01848*.

Wang, Y.; and Blei, D. M. 2019. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528): 1574–1596.

Wen, Z.; and Yin, W. 2013. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1): 397–434.

Worrall, D. E.; Garbin, S. J.; Turmukhambetov, D.; and Brostow, G. J. 2017. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5028–5037.

Zhang, K.; Zhang, J.; Huang, B.; Schölkopf, B.; and Glymour, C. 2016. On the Identifiability and Estimation of Functional Causal Models in the Presence of Outcome-Dependent Selection. In *UAI*.