On Computing Counterfactuals for Causal Fairness

by

Ayan Majumdar

Master's Thesis

Networked Systems Chair
Max Planck Institute for Software Systems

Faculty of Mathematics and Computer Science
Department of Computer Science
Saarland University

Advisor
Prof. Dr. Krishna P. Gummadi

Reviewers
Prof. Dr. Krishna P. Gummadi
Prof. Dr. Isabel Valera

March, 2021

Max
Planck
Institute
for
Software Systems

UNIVERSITÄT
DES
SAARLANDES

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis.

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken, _____ _____

(Datum/Date) (Unterschrift/Signature)

Dream, dream, dream! Conduct these dreams into thought, and then transform them into action.

— *Dr. A.P.J. Abdul Kalam*

Abstract

Causal reasoning has become a crucial tool for analyzing fairness in algorithmic decision-making systems. One of the most popular notions of causal fairness is that of counterfactual fairness. It defines a decision-making system to be fair if an individual and its counterfactual are provided the same outcome. However, this analysis crucially relies on the faithful computation of the counterfactual features for the given individual. Traditionally, causal methods have been used for the purpose. Unfortunately, these methods need strict assumptions about the underlying causal process that generates the data. Any mismatch between the assumptions and the true process often leads to significant errors in counterfactual approximations. This is especially problematic in the complex societal setting of fairness studies, rendering these causal methods impractical for the same. In this work, we aim to reduce some of these assumptions to achieve practical operationalization of counterfactual fairness. We do so by highlighting the primary considerations implicit in most fairness scenarios. We demonstrate that these main assumptions alone are sufficient for our goal, and we can reduce the supplementary causal considerations. Through rigorous empirical evaluation, we show that we can use deep generative models such as variational autoencoders to faithfully approximate the counterfactual mappings without extensive causal assumptions. We exhibit how these approximated counterfactuals could be used as a practical fairness tool, especially to audit downstream predictive systems. Finally, we highlight how the deep generative models could be used to train predictive systems that satisfy the notion of counterfactual fairness despite working on fewer assumptions.

Acknowledgements

This work would not have been possible without the support of a great many people. I would first like to express my deepest gratitude towards my advisor Prof. Dr. Krishna P. Gummadi, who provided me with the opportunity to work on this fascinating research topic. Your guidance and support was central to the successful completion of this work. I would like to thank Prof. Dr. Isabel Valera for reviewing and providing valuable insights on the technical aspects of the work. The formulation of the methodology and experiments would not have been possible without your significant support.

I would also like to thank Preethi Lahoti for being an incredible mentor, providing constant support and guidance in my research. I would like to thank Junaid Ali and Till Speicher who helped and guided me through my initial days of research on the topic. I want to additionally acknowledge my colleagues at the Max Planck Institute for Software Systems for engaging discussions about my work that helped me get important feedback for improvement.

I would like to thank my mother, my father and my brother for their constant love, support, and encouragement. You helped me push myself ever forward, even through the toughest of times. This thesis would not have been possible without the love and support of my dearest Debasmita Lohar. Your help in writing this thesis was invaluable, while your endearing love and support helped me through the many challenges over the past year. Finally, I would like to thank my friends in Saarbrücken and back in Kolkata for their encouragement. This thesis would not have been possible without you all.

Content

Nomenclature

List of Abbreviations

CF	Counterfactual
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
conf.	Confounder
CVAE	Conditional variational autoencoder
DAG	Directed acyclic graph
ELBO	Evidence lower bound
exog.	Exogenous noise
FYA	First Year Average
GAN	Generative adversarial network
KL	Kullback-Leibler divergence
LSAT	Law School Admission Test
MCMC	Markov chain Monte Carlo
MMD	Maximum mean discrepancy
RMSE	Root mean square error
SCM	Structural causal model
SGD	Stochastic gradient descent
UGPA	Undergraduate Grade Point Average
VAE	Variational autoencoder

List of Symbols

$\delta(\cdot)$	Dirac-delta probability distribution
ϵ	Random exogenous noise
$\hat{X}$	Estimated feature value
$\hat{Y}$	Predicted label
λ	Hyperparameter for optimal transport (FlipTest)
$\leftarrow$	Intervention
$\mathbb{D}_{KL}$	Kullback-Leibler divergence
$\mathbb{E}$	Expectation
$\mathcal{M}$	Causal MCMC models
$\mathcal{N}$	Gaussian distribution
μ	Mean of Gaussian
σ	Standard deviation of Gaussian
$\text{do}(\cdot)$	Do-operation for intervention
A	Sensitive features/attributes
b, c, d, w	Parameters of structural equations
$c(\cdot)$	Optimal transport cost
D	Discriminator of GAN
F	Factual data instance
f	Structural equations
f'	Intervened structural equations
G	Generator of GAN
$L, \mathcal{L}$	Loss functions

N, n	Total number of samples
$P(\cdot), p(\cdot)$	Probability
p_θ	Decoder of VAE
Q, K, U_J, U_D	Confounding variables
q_ϕ	Encoder of VAE
U	Hidden causal factors
X, V, E	Observed features
X^c	Counterfactual data features
X^f	Factual data features
Y	True outcome label
Z	Latent distribution space
z	Latent space random sample

Dedicated to my family and my loved one.

Chapter 1

Introduction

In recent years, we have seen wide-spread application of machine learning models in various fields like image processing, linguistics, search, and recommendation. In image processing, these models not only match humans but also surpass them in classifying images [1] and recognizing faces from images [2]. These models can also automatically generate text for e-mails (Google Smart Compose) [3], provide smarter and more personalized recommendations (LinkedIn Talent Search) [4, 5]. This has mostly become possible because of easy access to large amounts of data and recent developments in complex algorithms to process them. Owing to such success, people have started using machine learning models for decision-making in more societal situations. This includes high-stakes critical scenarios such as recidivism prediction [6], credit scoring [7] and healthcare [8]. Machine learning has enabled large-scale automation of complex decision-making processes, while maintaining high accuracy in its predictions.

Despite the rapid deployment in social settings, the actual effects of these automated decision-making systems on the society are not being assessed [9]. This lack of analysis might lead to potential bias and discrimination in the predictions made by such algorithms with respect to some *sensitive features* like race, sex or age. In fact, recent studies *have* discovered potential bias in various applications such as facial recognition [10], text generation [11] and recommendation systems [12]. Worrying problems of bias and discrimination have also been exposed in the systems deployed in critical decision-making situations. For instance, the recidivism prediction system named Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) was found to be biased against African-American inmates [13]. A similar study on algorithmic systems in healthcare [14] found that a predictive model widely used to identify patients with complex health needs consistently gave Black patients lower risk scores, leading them to get less treatment. Likewise, a model trained to predict the credit risk of individuals was found to be biased against certain races [15]. There is, therefore, a clear need to

analyze and study these systems not only with respect to *accuracy*, but also regarding how they *treat* people belonging to different sensitive groups.

This necessity led to multiple works on operationalizing different notions of bias and fairness for machine learning systems [15–17]. These works primarily focus on (i) encoding the various fairness notions from different domains (including the law) into mathematical formulation, (ii) and finding methods to optimize for these formulations. This operationalization allowed designing decision-making systems that satisfy certain statistical notions of fairness with respect to the outcome across different sensitive groups (race or gender). However, such bias may be deeply encoded in the data we collect due to historical discrimination in our society. Let us consider *redlining* for instance. In the United States, the systemic denial of housing in developed neighborhoods prior to 1968 has today led to significant gap in wealth and opportunities for minority groups. Any data collected in these circumstances for future predictions would be inherently biased. The early works on algorithmic fairness did not consider the *causes* of bias in the data. As a result, it is quite counter-intuitive to understand the actual source of bias in the system, and which fairness notion should be used to mitigate the same [18].

Recently, there has been a growing interest to address these shortcomings in fairness studies by using causal reasoning and analysis [19–22]. With the help of causal reasoning regarding all features in the data, it is possible to analyze and reason about the potential sources of bias in the decision-making system. Consequently, multiple definitions of causal fairness have been proposed, the first and most popular of which is the individual notion of *counterfactual fairness* [19]. This fairness notion states that "a decision is fair if it is the same in the *actual world* and a *counterfactual world*". Central to such a notion is the ability to compute counterfactuals for a given individual data point in the actual world. In the context of algorithmic fairness, this computation typically boils down to being able to *intervene* on (modify) the protected features[1] (e.g., sex, race) of individual data points in order to generate their corresponding counterfactual quantities [23]. This allows us to map any individual data point to its corresponding *counterfactual* point in a different sensitive group. This helps in analyzing the fairness of any decision-making system by computing the difference in the predictions provided to the actual and counterfactual data points.

In general, counterfactual fairness analysis follows a model-first approach using causal methods from the causality literature [23, 24]. These methods require complete causal knowledge about the process that generates the data, which constitutes: (i) the cause-effect relationships among *all* the observed features of the data, along with any hidden

[1]Concerns have been raised with regards to the inclusion of and intervention on sensitive features in causal reasoning. We highlight these concerns in Section 2.8.

factors that may be present in the generative process. These are specified using a full *causal graph* (ii) the exact quantitative dependencies among the features and the hidden factors, as defined by corresponding *structural equations*. This explicit knowledge about the data generative process allows for the computation of *counterfactuals* by performing the abduction-action-prediction steps [23]. However, complete causal knowledge is a very strong assumption as obtaining it is impossibly hard in most practical settings. This is true especially for societal situations, where it is very difficult to know about the exact quantification of the complex interactions between different features. Thus, to generate counterfactuals, traditional causal methods might end up making assumptions about the data generative process that is significantly different from the true process. This incongruity in the assumptions might even lead to severe errors in their counterfactual estimations. Hence, although appealing, the applicability of counterfactual fairness has been limited in practice.

In this thesis, we aim to reduce the amount of causal assumptions on the data that are usually required by traditional causal methods for fairness studies. In the existing literature studying *counterfactual fairness* [19–22], we have observed the prevalence of certain implicit structures: (i) observed features are the result of an individual's membership in a sensitive group along with some hidden/exogenous factors, and (ii) all individuals are assumed to have the same distribution for these hidden factors. These are the *main* assumptions that are necessary about the data generative process to study *counterfactual fairness*. As we show, the further assumptions regarding the cause-effect relations amongst the other features, and the quantifying equations are *not required*.

One potential concern remains: the reduction of these causal assumptions may lead to inaccurate counterfactual computations. In this work, we also show that reducing the supplementary causal considerations does not affect the approximation quality of counterfactual quantities for fairness purposes. Computation of counterfactuals requires knowledge of the data generative process. Traditional causal methods use additional strict assumptions for the purpose. We demonstrate that we can instead utilize deep generative models like the conditional variational autoencoder (CVAE) [25, 26] to *approximate* the generative process. This estimation enables the production of accurate counterfactual mappings without such additional causal assumptions.

For validation of our hypothesis and assumptions, we conduct a rigorous experimental evaluation using multiple synthetic and real-world semi-synthetic setups. By comparing with actual ground-truth counterfactuals, we show that the CVAE, while working with fewer assumptions, provides accurate approximations of the counterfactual mappings across different settings. At the same time, we highlight that traditional causal methods can provide accurate counterfactuals only when they have *perfect* knowledge

of the true process (*oracle models*). Any inconsistencies between the strict assumptions they make and the true underlying process lead them to make significant errors in the counterfactual approximation. We demonstrate that the CVAE is extremely flexible in modeling the generative process across multiple scenarios and perform faithful counterfactual estimations. Hence, the CVAE provides us with an easy and practical approach to operationalize the notion of counterfactual fairness.

Contributions To summarize, this thesis makes the following contributions.

1. We present the core assumptions about the data generative process made in the counterfactual fairness literature, and show how we can reduce some additional causal assumptions made by traditional methods.
2. Through rigorous comparisons, we show that deep generative models like the CVAE can be used as a practical tool to accurately approximate counterfactuals while working on the reduced assumptions.
3. Through a fairness auditing case-study, we demonstrate that the approximated counterfactuals are not only accurate, but also useful for practical fairness analyses.
4. We highlight that the CVAE model also allows us to train predictive systems that satisfy the notion of counterfactual fairness; something that traditionally required strict assumptions.

Thesis structure The rest of the thesis is organized as follows. We give a brief overview of the different background topics that are necessary for the understanding of our main contributions and evaluations in Chapter 2. In Chapter 3, we discuss in detail our approach that leads to the practical operationalization of counterfactual fairness using CVAE models. We provide a brief survey of the related literature in the space of causal reasoning, fairness, and generative models in Chapter 4. In Chapter 5, we explain the experimental setups, demonstrating the findings of our evaluations, along with a discussion of them. Finally, in Chapter 6, we conclude our observations and discuss the limitations and potential directions to pursue as future work.

Chapter 2

Background

This chapter aims to introduce the major concepts that are crucial to the understanding of our approach and methodology. First, we discuss the preliminary works on the fairness of algorithmic decision-making systems in Section 2.1. We introduce the notions of causal models in Section 2.2, interventions in Section 2.3 and computing causal counterfactuals in Section 2.4. Newer notions of fairness that use causal reasoning are introduced in Section 2.5. A traditional approach to approximating causal processes using Markov chain Monte Carlo (MCMC) models is introduced in Section 2.6. We introduce a class of deep learning models that are central to our approach: deep generative models in Section 2.7. We provide a discussion on interventions with respect to sensitive attributes in Section 2.8. Finally, a discussion on legal discrimination and application of counterfactual reasoning is presented in Section 2.9.

2.1 Algorithmic fairness

With machine learning based predictive models becoming more and more ubiquitous, there has been a lot of work aimed at studying their fairness implications. These fairness works usually work on classification models with respect to some sensitive features A (race, sex) to equalize certain outcome statistics. The notions are defined either on a **group-level** with respect to A, or at **individual-level**. We discuss some of the most popular fairness notions here. We denote the observed features as X, true output label as Y, predicted label as $\hat{Y}$.

1. Fairness through unawareness *A decision-making system is fair if it does not explicitly use any sensitive features in its process* [27]. This definition of fairness actually satisfies the legal requirement for disparate treatment. Unfortunately, there can still be potential bias owing to hidden correlations between the sensitive attributes and

other features X. For example, consider the scenario of predicting credit scores, where salary or net income is an important feature. Owing to historical discrimination, female individuals have significantly lower salaries in the workforce [28]. Hence, a system that does not use "sex" as a feature can still be biased against females.

2. Demographic parity *A decision-making system whose outcome is independent of the sensitive features satisfies this notion* [29]. So, for a predictor to satisfy demographic parity, $P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$. This notion satisfies the legal definition of disparate impact. However, this notion by itself also has some shortcomings. For instance, a predictive model that randomly gives the positive outcome to male and female individuals would satisfy this notion of fairness. But clearly, this would end up selecting plenty of non-deserving candidates.

3. Individual fairness *A decision-making system satisfies the notion of individual fairness if it provides similar outcomes to similar individuals* [17]. This notion is clearly a stronger notion of fairness as it can potentially become fair at the level of each individual. For operationalizing this notion, we require access to a well-defined distance metric $d(i, j)$ in the data-feature space. This metric allows measuring the similarity between different individuals, and using this we can have data points that have similar features $X_i \approx X_j$ also getting similar decisions. However, the definition of the distance measure is crucial here and can be tricky to explicitly define it mathematically.

These notions of fairness were successful in operationalizing different fairness definitions for algorithmic systems. But they also have some shortcomings. For example, these definitions fail to highlight the source of historical bias that might exist in the data. Similarly, it is not quite obvious how to tackle such bias, which often makes fairness studies complicated. Integrating causal reasoning [23] in the study of fairness helps in understanding the source of bias and solving for it as well.

2.2 Structural causal models

Causal reasoning makes use of causal methods, that rely on the complete definition of the underlying assumptions regarding the true data generative process. This is generally specified with the help of a structural causal model or SCM [23]. The SCM helps to describe all relevant observed features X and hidden factors U that are involved in the data generation. It also defines how these different features interact with each other. So, causal methods can have complete knowledge of the generative process through its

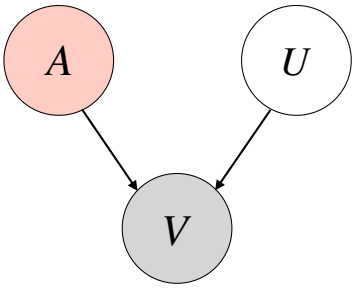


Figure 2.1: A simple illustrative causal graph. V, A denote observed variables, whereas U denotes the unseen noise.

corresponding SCM. This SCM consists of the following components, a causal graph and a set of structural equations.

1. Causal graph As discussed in [23], each SCM can be associated with a graphical causal model, or more simply, a *causal graph*. This is represented by a directed acyclic graph (DAG) containing nodes that represent both the observed and hidden variables. The edges in the graph represent which variables are related to each other, while the directionality of the edge represents the cause-effect nature of the relationship. Figure 2.1 represents a simple causal graph, where V, A denote observed and U denotes hidden variables. Note how the directed edges define that V is caused by A and U.

2. Structural equations An SCM is also defined by a set of equations f that helps in determining the value of each variable in the model with respect to the other variables. This set f is collectively termed the "structural equations". These equations define the exact realization of the variables in the model. Each observed feature X would be realized by its corresponding function f_X as: $X = f_X(\text{Pa}(X), U)$; where $\text{Pa}(X)$ represent the values of the parents of the node X in the graph. For our simple graph in Figure 2.1, we will have $f := \{f_V\}$, where $V = f_V(A, U)$. The form of the equations in f can be anything ranging from additive linear to complex non-linear relationships.

Rule of product decomposition The use of DAGs as causal graphs can help to understand the causal data process efficiently using the rule of product decomposition. This allows us to denote the joint distribution over the data $P(X)$ as follows:

$$P(X) = \prod_i P(X_i | Pa(X_i)) \tag{2.1}$$

For the model defined by Figure 2.1, we have $P(V, A, U) = P(U)P(A)P(V|A, U)$.

2.3 Causal interventions

Interventions are an extremely important tool in causal studies. Intervening on specific variables allows us to eventually predict the outcome caused by the same. Whenever possible, randomized controlled experiments are performed to study intervention effects of specific features. For example, a drug company might perform a randomized controlled experiment to study the effect of an intervention with respect to providing a particular medicine to see the outcome in patients. At the same time, SCMs allow studying such interventions more easily as hypothetical operations. In SCMs, interventions usually involve surgically modifying the causal graph, by forcing the variable of interest to a specific value, for example, $A \leftarrow a$. We can also denote this by the *do-operation* $\text{do}(A = a)$. Note the difference between a conditional and interventional distribution [23]. Whereas $P(V|A = a)$ represents the distribution of V among individuals who have $A = a$, $P(V|do(A = a))$ represents the distribution of V if every individual in the population were fixed to have $A = a$. Later on, we also represent interventions interchangeably with $P(V_{A \leftarrow a})$. The probabilities can be represented by the corresponding structural equations f whenever possible.

Adjustment Formula Interventions can help in the estimation of causal effects in a particular model. The adjustment formula helps in this purpose by *adjusting/controlling for* variables that are of interest. For our simple example, to understand the effect of A on V, with the help of intervention on A and Equation 2.1 we have:

$$P(V|do(A = a)) = \int f_V(V|A = a, U)P(U)dU \tag{2.2}$$

2.4 Causal counterfactuals

Complete knowledge of the entire SCM is a very strong assumption for most practical scenarios. However, for causal analysis, it allows for the computation of counterfactual entities. Let us consider our simple example from Figure 2.1. Counterfactuals represent scenarios such as: "the value of feature V had A been a'". Such quantities can be computed by following the 3 steps given by Pearl [23], along with Equations 2.1 and 2.2:

1. **Abduction**: Given an observed instance $V = v, A = a$, use the SCM to estimate posterior $P\left(U|V = v, A = a\right)$.
2. **Action**: Perform *intervention* on A by forcing $A \leftarrow a'$ in the SCM.

3. **Prediction**: Use estimated posterior on U and intervened equation f'_V to get counterfactual distribution[1]as:

$$P(V_{A\leftarrow a'}(U)|V=v,A=a)=\int f'_V\left(A\leftarrow a',U\right)P\left(U|V=v,A=a\right)dU \quad (2.3)$$

2.5 Causality and fairness

There has been great interest in applying the notions of causality to address the issue of fairness in algorithmic decision-making. This is generally studied with respect to some sensitive features A such as sex or race and how it affects observed features X and consequently, a potential decision Y. Inevitably, various definitions were developed. A fundamental notion of fairness with respect to counterfactual entities was defined in [19]. Fairness definitions with respect to the direct or indirect effects of A on the outcome were discussed in [30]. An even more fine-grained viewpoint to estimate the bias along specific paths in the causal graph was provided in [21]. Further discussions on studying discrimination through proxy and resolving variables are conducted in [31].

Counterfactual Fairness In this work, we focus on the individual notion of *counterfactual fairness* [19]. It states that a decision-making system is fair if an individual is provided the same decision in the *actual* world and a *counterfactual* world where they belong to a different demographic group. Formally, it is defined as:

$$P\left(\hat{Y}_{A\leftarrow a}(U)=y\mid X=x,A=a\right)=P\left(\hat{Y}_{A\leftarrow a'}(U)=y\mid X=x,A=a\right) \quad (2.4)$$

This estimates the total bias caused by A with no resolving variables. So for any particular individual, A should not be a cause for the outcome Y. A counterfactual for such an analysis is computed via an intervention on $A\leftarrow a'$, given the individual's hidden factors being U. These interventions are infeasible in reality and are thus generally considered as hypothetical computations. Such counterfactual quantities could be used to audit existing decision-making systems for fairness as defined by Equation 2.4. The framework also allows us to model a system that can be fair with respect to the definition using the hidden U, as discussed in [19].

[1]For a deterministic setup where X and U are bijective, we can calculate the counterfactual directly using $V=f'_V\left(A\leftarrow a',U\right)$. The estimation of U in **Abduction** would also be deterministic as $\delta(U|V=v,A=a)$.

2.6 Causal MCMC Model

Fairness studies through causal inference using traditional methods generally require strict assumptions about the causal structure and the equations pertaining to the data. This is because computing counterfactuals requires the estimation of the hidden factors U. However, even with perfect causal knowledge, the relationship between X and U may not be bijective. Hence, we may not be able to exactly estimate the hidden factors from the data. Thus, causal inference methods generally use probabilistic inference techniques. As seen in [19, 20], one may use probabilistic models like Markov chain Monte Carlo (MCMC) for causal analysis. Given the perfect knowledge of the causal structure of the data, the goal of these models would be to estimate a posterior distribution on $P_{\mathcal{M}}(U|X, A)$, where $\mathcal{M}$ denotes the MCMC model we use. As mentioned, the MCMC model would be heavily reliant on the strict causal assumptions it makes regarding the true data generative process. MCMC makes use of two properties [32] for inference:

1. **Monte Carlo:** The properties of a distribution can be estimated by examining random samples that are drawn from the distribution.
2. **Markov chain:** It involves generating a sequence of random samples where the current sample is dependent on the sample from the previous step.

MCMC methods combine these two properties to provide an inference technique that can sample from high-dimensional probability distributions. Given the data, it would learn a posterior on the hidden noise factors that allows us to draw useful samples of $U_i \sim P_{\mathcal{M}}(U|X_i, A_i)$ for given observed data-points. As we have seen in Section 2.4, this is the crucial first step in the estimation of counterfactuals using Pearl's method.

2.7 Deep Generative Models

Deep learning has made modeling high-dimensional data distributions for learning the generative process highly effective through the advent of deep generative models. These models allow a simpler approach to modeling complex data processes using deep neural networks. This simplification has also made these models useful tools in causal analysis and fairness [33, 34]. Here, we discuss two types of such models: the generative adversarial network (GAN) and the variational autoencoder (VAE).

2.7.1 Generative Adversarial Networks

This class of models aims to capture the generative process of a particular dataset using adversarial training [35]. It aims to train two separate deep models simultaneously:

- **Generator:** The generator (G) is a deep neural network that aims to capture the complex high-dimensional data distribution space, such that it can generate realistic data samples.
- **Discriminator:** This is another deep neural network, D, that aims to predict whether a given sample came from the original training distribution or from G.

In the training process, the two models play a min-max game that can be characterized as follows:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}\,(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.5}$$

where V represents the value function for the game, p_{data} represents the data distribution, $p(z)$ represents some random noise distribution. Eventually, G would be able to generate samples that are extremely realistic and mimics the actual training distribution.

FlipTest In the space of counterfactual mappings for fairness, GANs were used in FlipTest [33]. It aimed at approximating the mapping process through optimal transport [36]. By using optimal transport, FlipTest aims at finding the most optimal mapping of feature values between individuals of two groups. FlipTest uses GAN models to model the optimal transport process. In the process, this method manages to generate counterfactual mappings without strict causal assumptions. Instead of using random noise z as input to the generator, FlipTest uses data points from one sensitive group to generate an optimally mapped data point in another group. It uses a variant of Wasserstein GAN [37] with the generator loss function:

$$L_G = \frac{1}{n}\sum_{x\in X} D(G(x)) + \frac{\lambda}{n}\sum_{x\in X} c(x, G(x)) \tag{2.6}$$

Here, $c(\cdot)$ represents the optimal transport cost, that is balanced with the traditional generator loss via the hyperparameter λ.

2.7.2 Variational Autoencoders

This is a class of latent-variable generative models that aims to capture the joint data distribution through a latent distribution space. As we show in Chapter 3, we utilize a

variant of this model to learn complex data processes to generate counterfactuals. The VAE consists of two separate deep models:

- **Encoder** (q_ϕ): This model encodes the high-dimensional data distribution in the latent space. It learns the parameters of the latent distribution space from the observed data.
- **Decoder** (p_θ): This model aims to regenerate realistic data samples from the encoded latent space. Given a sample from the latent space, it generates a realistic data point.

The latent Z is usually considered to match a simple prior distribution, for example, a standard normal $Z \sim \mathcal{N}(0, 1)$. The two models are trained together end-to-end by maximizing what is known as the evidence lower-bound objective (ELBO) [38]:

$$\log p_\theta(X) \geq \mathbb{E}_{q_\phi(Z|X)} \left[\log p_\theta(X|Z)\right] - \mathbb{D}_{KL} \left[q_\phi(Z|X) \,||\, p(Z)\right] \tag{2.7}$$

The data-evidence $p(X)$ is maximized via the lower-bound approximation. This involves the decoder that maximizes the data log-likelihood given the latent distribution (*first term*) and the encoder that minimizes the Kullback-Leibler (KL) divergence $\mathbb{D}_{KL}$ between the estimated posterior and the prior on Z (*second term*). The models are trained using backpropagation stochastic gradient descent (SGD). However, these models have stochastic components, and need special care to allow for backpropagation.

Reparameterization trick In order to train the VAE model, we need to sample $z \sim q_\phi(Z|X)$ from the learnt posterior. Unfortunately, SGD using backpropagation cannot handle stochastic units in the networks. On the other hand, it can handle stochastic inputs. Therefore, the solution is to move this sampling process to the input, what is termed as the "reparameterization trick" [38]. This is especially simple for Gaussian priors. Consider the learnt posterior $q_\phi(Z|X)$ is a Gaussian $\mathcal{N}(\mu(X), \sigma(X))$ where the parameters of the Gaussian are learnt by the encoder from data X. The idea is simple: instead of sampling the latent z from the posterior distribution directly, sample a random noise ϵ that acts as a stochastic input instead. Then, sampling z from this posterior is the same as having $z = \mu(X) + \epsilon * \sigma(X)$, where $\epsilon \sim \mathcal{N}(0, 1)$. This allows backpropagation to be used to train the models.

2.8 Intervention of sensitive attributes

There has been extensive discussion on the inclusion of and intervention on sensitive attributes in a causal graph. We discuss some of them in this section.

Lack of support for constructivist view Traditional studies on causality and fairness have often considered sensitive attributes like race and gender from a naturalist viewpoint. This view argues that such attributes are indeed biological, and hence are intrinsic to individuals. In fact, works as [19–22] have made assumptions that support this view: sensitive features are root nodes in any causal graph. However, there is another conflicting viewpoint, constructivism, that considers race and gender as social constructs. According to them, the individuals in a society are classified into one of these groups owing to how they are perceived by that society. The shortcomings of analyzing causal effects of attributes like race, especially from a constructivist point of view, are discussed in [39]. It mentions that the issue with causal and counterfactual reasoning with race is that we usually reduce race to its biological signs. However, it fails to satisfy the *constructivist* view of race with which many scholars agree. Similar implications are further studied for race in [40]. It mentions that race can be used in causal studies, however, it needs to be operationalized as a "bundle of sticks". It highlights some problems of modeling race as a single entity. For example, in such settings, most factors that social scientists control for like education or neighborhood, become post-treatment factors, and thus introduce bias in the causal analysis. This is a common problem known as *post-treatment bias*. Instead, it motivates operationalizing race with a constructivist view, as a bundle of many different factors, some mutable and some immutable. A similar discussion about the ontology of sex in causal graphs and how such *sex groups* need to be encoded can be found in [41].

No causation without manipulation The notion of potential exposability is used in [42] to argue against using features like race in causal studies owing to "no causation without manipulation". It explains that for causal inference, it must be possible for each unit of the population to be exposed to any of the "causes" in a controlled experiment. Sensitive features like race or sex are intrinsic. So they do not satisfy this notion, as modifying it would lead to "changing the unit in some way such that it no longer remains the same". There have been several counterarguments regarding this as well. Race or sex can be considered as causes if "events" that led to them, for example, conception, are considered in the causal reasoning [43, 44]. The definition of "cause" is broadened to include extrinsic and intrinsic ones in [45]. It notes how "a synthesis of intrinsic and extrinsic" factors can provide a more adequate causal picture.

In extension to previous fairness studies, our assumptions only satisfy the naturalist view and not the constructivist view. We note that this requires further analysis and discussion. We also follow the latter notion and consider that attributes such as race and sex can be causes and intervened upon *hypothetically* for causal analysis of fairness. We consider race or sex to be random variables and individual nodes in a causal graph, following most fairness studies [19–22].

2.9 Counterfactuals and legal discrimination

Causal analysis and using counterfactuals have gained popularity in fairness studies related to algorithmic decision-making systems. However, there remains a gap between the developed notions and legal definitions of discrimination. There have been limited discussions in the legal domain regarding causal analysis. The study conducted in [46] shows how counterfactuals are generally used in legal disputes, how they might be interpreted and how they should be handled in a meaningful way. It highlights some of the challenges and provides some arguments for when counterfactuals might be useful and how they may be interpreted in the legal domain. For example, to ask the right questions, the cause should be "stated in legally relevant terms" whereas the effect should also be within "the perimeters of legal relevance". At the same time, a lengthy discussion regarding the presence of causal and counterfactual reasoning in anti-discrimination cases is given in [47]. It showcases how the inquiry process in courts could be mapped to causality – *status inference* (plaintiff highlights how race or gender is a possible explanation of the contested decision); *neutral explanation* (defendant attempts to explain the same decision with respect to "status-neutral" terms); *causal attribution* (the court analyzes whether the contested decision was made due to factors related to status influences or neutral explanations). Nonetheless, formal causal analysis and causal requirements for discrimination is still contended in the law. However, as discussed in [48], there have been some *actual* legal scenarios where counterfactual interpretations have been made in anti-discrimination cases by lawmakers and legal representatives. The ruling in *Carson vs Bethlehem Steel Corp.*, 70 FEP Cases 921, 7th Cir. (1996) had the lawmakers remark:

> "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin, etc.) and everything else had been the same."

The recent Harvard admissions lawsuit (*Students for Fair Admissions, Inc. v. President and Fellows of Harvard College et al.* Civil Action No. 14-14176-ADB (D. Mass)) had a plaintiff's expert noting:

> "Consider the example of an Asian-American applicant who is male, is not disadvantaged, and has other characteristics that result in a 25% chance of admission. Simply changing the race of the applicant to white—and leaving all his other characteristics the same—would increase his chance of admission to 36%"

These cases clearly show clear examples of counterfactual reasoning being used in legal statements. However, it should be noted that these situations are quite unique, presented in only a few cases, and have limited precedence. One interesting thing to note is the common assumption of "keeping everything else" the same. This often refers to simply changing the sensitive features while keeping all other data features the same. A direct causal interpretation means that sensitive features would not *cause* any features in the data. This, of course, ignores the societal causal effects that race or sex can have on other factors due to historical discrimination and systemic racism. These discussions highlight how formal counterfactual reasoning could become an important tool for evaluating fairness and bias in law, but one that needs significant discussion and analysis.

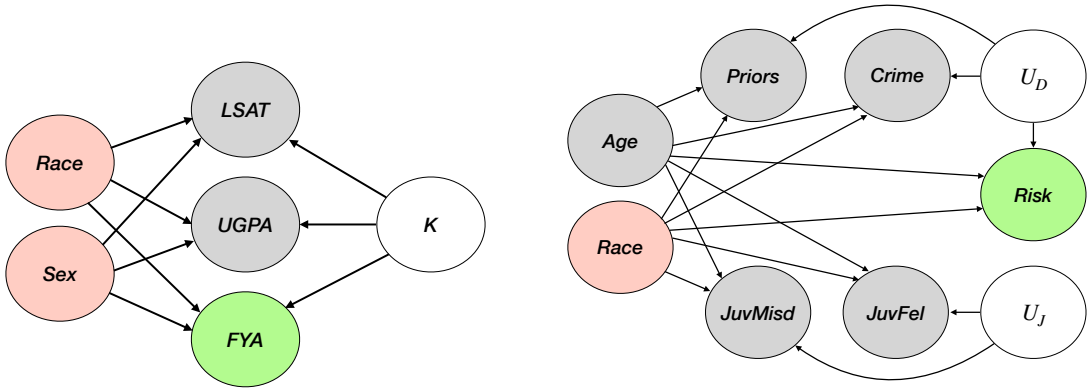
Chapter 3

Approach

A direct approach to analyzing any decision-making system for *counterfactual fairness* would be to ask: "Would a datapoint x get similar outcome if it belonged to a different sensitive group?". This would require calculating the counterfactual of x and then obtaining the updated outcome from the decision-making system for the same. However, computing such counterfactual entities is a challenge. Traditional causal approaches require strict assumptions about the SCM and the generative process, as discussed in Section 2.4, that make them impractical. In this chapter, we show how we can reduce some of the assumptions and operationalize counterfactual fairness. We highlight the common implicit structures that are relevant for fairness studies in Section 3.1. These allow us to reduce the additional assumptions and leverage deep generative models to learn the generative process with these fewer assumptions (Section 3.2). Finally, we delineate how we could use such models to approximate counterfactual entities paralleling Pearl's steps (Section 3.3). For further discussion, we introduce two illustrative examples here.

Illustrative Example 1 In the law school admission scenario [49], students complete their undergraduate degrees, take a standardized test and apply to law schools. There are three types of variables: (i) observed sensitive (A) such as sex and race, (ii) non-protected observed features (X) such as undergraduate grade-point average (UGPA) and Law School Admission Test (LSAT) test score, (iii) hidden (U) such as intrinsic knowledge of a student. Potential candidates are selected by trying to predict the First year average (FYA) they would have scored in Law school *were they to be admitted*. Due to historical discrimination, race and sex causally affect the test scores. Figure 3.1a describes the causal process as per [19].

Illustrative Example 2 In the United States, an algorithmic system named the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is



(a) Law school admissions (b) COMPAS recidivism risk

Figure 3.1: Illustrative causal graphs for (a) Law school admissions and (b) COMPAS recidivism risk. (a) Race and sex are sensitive; LSAT and GPA scores are observed features. Knowledge K is the confounder. (b) Race is sensitive; age, juvenile felonies and misdemeanor, priors, crime are observed features. Criminality is the hidden confounder (U_J, U_D). Note that in both cases, exogenous noise variables for each observed variable are not shown, but present in the process.

used to predict the recidivism risk for criminal defendants [6]. A record is kept of the defendants' age, crime, prior counts, juvenile felonies and juvenile misdemeanours. These are the observed features X. The sensitive feature A is considered the race of any individual. The hidden variables in this case are considered the intrinsic criminality of any defendant U_J, U_D. The outcome variable is the recidivism risk score for the defendant. Due to historical reasons, all the criminal features are causally dependent on race as well as age. The representative causal graph for COMPAS recidivism is shown in Figure 3.1b as per [20].

3.1 Assumptions

Most studies conducted on causal analysis and fairness, in particular counterfactual fairness, encompass very similar assumptions regarding the data and the relations among the various features. As we would go on to show, these actually allow us to operationalize counterfactual fairness using more practical approaches. In this section, we highlight the common assumptions that encompass fairness studies. These are the main assumptions that we also make about the data generative process for our analyses regarding counterfactual fairness. As we discuss, these are in fact the primary and only assumptions we need to study counterfactual fairness, and can reduce the additional considerations made by causal methods. For all corresponding discussions, we use the following notions. All hidden noise factors involved in the causal generative process are grouped into a single notative variable U. All sensitive features that are present are

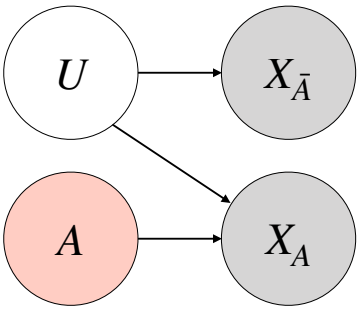


Figure 3.2: Assumed simplified causal graph that generalizes to many fairness scenarios. U represents the hidden and exogenous factors, A represents the observed sensitive attributes, e.g., race, sex, etc. All the observed non-sensitive variables are grouped in X. $X_{\bar{A}}$ are the observed variables which are only caused by the hidden exogenous factors, while X_A are the observed variables which are additionally caused by A.

concatenated into A, while the observed non-sensitive features in X. The assumptions are as follows:

(i) A are root nodes, not affected by any other variables. They may contribute to the generation of some features X, which we term X_A while the rest are $X_{\bar{A}}$.

(ii) U affects X independent of A, so, there can be no causal link between A and U.

These can be summarized by the simplified causal graph shown in Figure 3.2. As discussed in Section 2.8, for most fairness studies, sensitive features are considered to be intrinsic factors. Additionally, as per [19], the notion of *ancenstral closure for sensitive attributes* also needs to be valid. This mentions how sensitive features need to be the root nodes in a causal graph, and if any variables exist that are parents of A in the graph, then they also need to be subsumed in A. These notions validate the first assumption. Further, as per the second point, no direct interaction between A and hidden U is considered. This is implicitly assumed by the majority of fairness studies. Any (incorrect) assumption of confounding between A and U would mean, for example, "intrinsic knowledge" of a candidate depends on their race or sex, which is controversial. Additionally, traditional causal analysis also considers hidden variables U to be root nodes [23]. These validate the second assumption about the generative process. Interestingly, the second assumption regarding no causal links between A and U also leads to an interesting corollary property: "monotonicity" of feature values X across different groups of A. This can be explained by a simple example. Consider we are looking at the success of sports personalities, where A represents the sports each individual plays. Non-monotonicity across features might occur if, with respect to similar levels of success, basketball players require greater height while football players require shorter heights. This can happen if the exogenous factors that lead to the generation of height is causally

affected by A in some way. These situations are usually not tackled by causal studies or fairness studies, and are also not relevant for our analyses.

If we look at these assumptions from purely a causal perspective, they would seem extremely restrictive. However they are implicitly valid and widely prevalent in the literature that studies *counterfactual fairness* [19–22, 31]. For instance, the graph for studying fair law school admissions [19] (Figure 3.1a) *can be* represented in a more concise way through Figure 3.2. Note how race, sex can be grouped into A, hidden factors including knowledge into U and observed test scores into X_A. Similarly, the graph for fair recidivism risk prediction (Figure 3.1b) as in [20] can also be easily subsumed by Figure 3.2, with Age as $X_{\bar{A}}$, criminality as U, race as A and the rest of the observed features as X_A.

But even in these situations, traditional causal methods require additional knowledge of the complete causal structure and equation forms for *all the features*. This explicit knowledge allows these models to use Pearl's 3 steps [23] to generate counterfactuals. *For our purposes, we do not assume any further knowledge on the causal graph nor the structural equations that would determine the relationship for observed X in the causal graph.* The assumptions in this section are in fact the *main* considerations required for studying *counterfactual fairness*. All additional assumptions that are made by traditional methods can be reduced, thereby allowing for a more practical approach to analyzing this fairness notion. As we show later, we can work with these reduced assumptions and use a deep generative model, the conditional VAE (CVAE), to approximate the counterfactual generation process. To this end, we introduce 3 steps that parallel Pearl's steps in Section 3.3.

There can still be scenarios where graphs different to what we assume describe the data. In fact, the issue of different causal graphs resulting in the same joint data distribution is studied in [31]. But for studying *counterfactual fairness*, we do not need to worry about this owing to no consideration for path-specific or resolving effects. For counterfactual fairness, we need to intervene on A and estimate the causal effect of the hidden U to generate meaningful counterfactuals. This requires the estimation of the total causal effect of A, U on the observed features X. Note that we can estimate these effects from the data using only the primary assumptions explained in this section while reducing the further causal considerations. Thus, the assumptions and related structure of Figure 3.2 are sufficient for studying and practically analyzing *counterfactual fairness*.

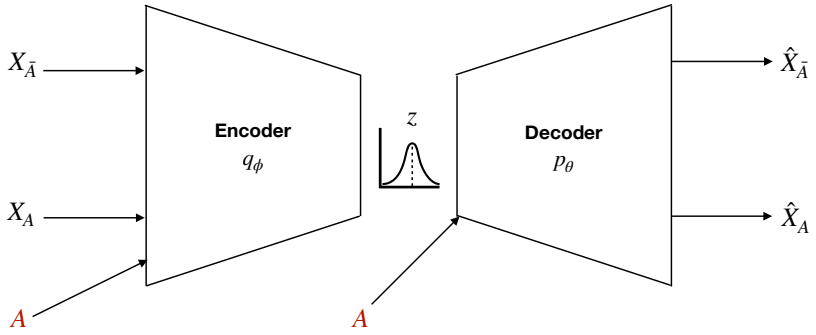


Figure 3.3: Conditional variational autoencoder (CVAE) model used for learning the data-generative process. We wish to intervene on sensitive features A for fairness analysis. So, we learn a conditional model with respect to A, which serves as additional inputs to the encoder and decoder models.

3.2 Learning latent exogenous factors

To study counterfactual fairness, the goal is to approximate the distribution of counterfactuals X^c as $P(X^c_{A\leftarrow a'}|X^f, A)$ where X^f is the given factual data, A is the sensitive feature where we apply the intervention $A \leftarrow a'$. For fairness scenarios working with the assumptions of Section 3.1, we would not need to estimate the exact values of each causal U. Indeed, it would be enough to estimate the total effect of U and A on the data process. To this end, we learn the conditional data distribution $p(X|A)$. In this setting, we would use latent-variable modeling to approximate this process. We capture the total effect of U on the data process through a latent space Z:

$$p(X|A) = \int p(X|Z, A)p(Z)dZ \tag{3.1}$$

In causality, the hidden variables U usually indicate some meaningful factors, e.g., knowledge of students, along with the noise variables affecting individual observed features. In this case, Z can be thought of as non-interpretable latent factors of the data process commonly used in machine learning studies [50]. As we show, for our scenarios, counterfactuals can be estimated through the latent Z via deep generative modeling.

To learn this generative process without access to the exact SCM, we rely on a particular type of deep generative models: the conditional variational autoencoder (CVAE) [25, 26]. The CVAE is a special variant of variational autoencoder (Section 2.7.2). Whereas the VAE models the joint data distribution through a latent-variable setup, the CVAE approximates the data process *conditioned* on some observed features. In our case, the conditional variable would be the sensitive feature A. Similar to the VAE, the CVAE also comprises of two deep neural networks: a decoder p_θ that captures the generation process of features X given Z and A, and an encoder q_ϕ that aims to approximate

the posterior distribution of Z from the observed data X and A. The parameters of the neural networks are θ, ϕ respectively. These parameters are estimated during the learning process along with inference of the posterior distribution. Given data X, as input to the encoder, the decoder model outputs a reconstructed approximation $\hat{X}$. The interesting feature of this model is the specific input of A into both the encoder as well as the decoder models, compared to traditional VAE models. A schematic of this model is shown in Figure 3.3.

The learning and inference process involves the optimization of a conditional variant of the standard ELBO loss for CVAE [38]:

$$\log p_\theta(X|A) \geq \mathbb{E}_{q_\phi(Z|X,A)} \left[\log p_\theta(X|Z,A)\right] - \mathbb{D}_{KL} \left[q_\phi(Z|X,A) \,||\, p(Z)\right] \tag{3.2}$$

We optimize the conditional evidence (left) through the approximated lower-bound (right). The first term on the right represents the expected log likelihood of the data, where the expectation is computed over all the estimated latent factors. Our goal is to maximize the conditional log-likelihood of the data from the learnt latent space. The second term acts as a regularizer, using KL divergence to force the estimated posterior, $q_\phi(Z|X,A)$, to be as close as possible to a known prior distribution $p(Z)$. As is common in literature [25, 26, 38], we consider the prior to be a very simple standard normal, $p(Z) := \mathcal{N}(0,1)$. Note how given any sensitive group membership, this makes the model learn the posterior on Z conditioned on A that is also invariant with A. This modeling is causally grounded given our assumptions in Section 3.1, where the hidden factors are assumed to have no association to A. Similar to the methodology of [51], we use deterministic decoders in our model to reduce the sources of stochasticity. With these considerations, the right-side of Equation 3.2 can be rewritten as follows:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{q_\phi(Z|X_i,A)} \left(\frac{|| \, X_i - \hat{X}_i \, ||^2}{\sigma^2} \right) + \mathbb{D}_{KL} \left[q_\phi(Z|X_i,A) \,||\, p(Z)\right] \tag{3.3}$$

where σ^2 is a hyper-parameter similar to [52]. To estimate the latent factors, we minimize the above loss function. Note how we can simply replace the negative log-likelihood of the conditional data-distribution with a squared error term. This is especially valid because in this study we consider continuous data features. Note that the CVAE models do not assume knowledge of the descendants or nondescendants of A ($X_A, X_{\bar{A}}$), but simply work with the joint data distribution. Incorporating such knowledge into the model is left as future work.

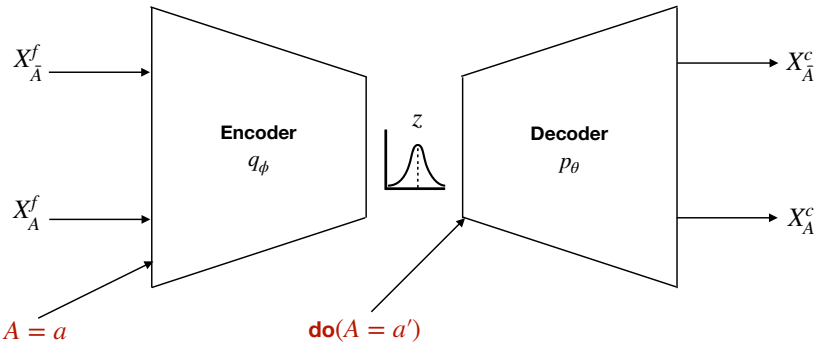


Figure 3.4: Conditional variational autoencoder (CVAE) model used for performing interventions with respect to A to approximate counterfactual data-points.

3.3 Approximating counterfactuals

In the previous section, we described how we can use a CVAE model to approximate the data distribution conditioned on the sensitive features A. As we work with the scenarios pertaining to *counterfactual fairness*, provided the assumptions described in Section 3.1, this allows the CVAE model to perform interventions and approximate counterfactuals. Although we do not measure each hidden causal U, we note that we do estimate the **total** effect of U and A on the data distribution. Hence, given a CVAE trained with our setup, we can actually approximate the counterfactual process. Although we cannot directly apply Pearl's 3 steps [23], we describe 3 steps that *parallel* the same. We assume, at test time, we are given some input *factual* data-point $F := \{X = x, A = a\}$. The steps that allow counterfactual approximation using CVAEs are as follows:

1. Given the trained encoder model and input data-point F, estimate the posterior distribution on the latent Z as $q_\phi(Z|F)$.
2. For the intervention step on $A \leftarrow a'$, explicitly set $A = a'$ at the input of the decoder model.
3. Given the posterior distribution on Z obtained from the encoder and the intervened A as input to the decoder, estimate counterfactual distribution on X^c using the trained decoder model:

$$P(X^c_{A \leftarrow a'}|F) = \int p_\theta \left(X^c|A = a', Z\right) q_\phi \left(Z|F\right) dZ \tag{3.4}$$

Even without exact inference of causal U, we are able to use the CVAE to estimate the total effect of U on X through the latent space Z. These steps show how this allows CVAEs to approximate counterfactuals for fairness purposes. Figure 3.4 visually describes the approximation process using the CVAE model. In practice, for deterministic

counterfactual estimation, we would need to perform Monte Carlo sampling from the posterior distribution $q_\phi(Z|F)$ as $Z^{(1)}, Z^{(2)}, Z^{(3)}, \cdots \sim q_\phi(Z|F)$. Thus, at the decoder, we can use such samples to estimate some samples from the counterfactual distribution. For a deterministic estimation, we can then simply compute the distribution mean from these Monte Carlo samples.

Issues of identifiability One recurring issue in causality is that of identifiability. As such, causal models generally assume "no hidden confounders" that affect both the cause and effect [23]. This specifies that in the assumed causal system, the cause and the effect variables both cannot be affected by an unmeasured confounding factor. Failing this, stricter assumptions about the process are required, or it is necessary to observe further features to mitigate the unmeasured confounding. As noted in Section 3.1, we assume that the U in the causal process would only affect X and not A. Likewise, U affects X independently of A. These are again assumptions that are usually true in many fairness-related causal studies [19, 34]. As noted earlier, although these seem restrictive from a causal viewpoint, they are widely used and valid for studying *counterfactual fairness*. Additionally, similar to [34, 50], we assume all confounders are represented well by some observed X. This denotes that unobserved confounding factors, if present in the generative process, are approximable from the observed data in some way. For example, let us consider for the law school setting, knowledge of individuals strongly affects the test scores. Therefore, given enough observed data with respect to the test scores of students, it can give us a strong enough signal to approximate the unseen knowledge as well. The approximation quality of the true data process by CVAE via Z is heavily dependent on the informativeness of X with respect to the unseen U. If this does not hold, the model might make inaccurate estimates. This inaccuracy might result in potential unfairness in the process, but the phenomenon would need to be further studied.

Chapter 4

Related Work

Our work aims to provide a practical operationalization for one of the most popular causal definitions of fairness – counterfactual fairness. We discussed our approach to the problem using deep generative models like the CVAE. In this chapter, we discuss several works that have been done in the related topics – *analysis of counterfactual fairness*, *deep generative models in causality*, *deep generative models in fairness*, *fair representation learning* as well as *counterfactuals for recourse and explanations*.

4.1 Analysis of counterfactual fairness

One of the first definitions of fairness that used causal analysis was *counterfactual fairness* [19]. The work motivated the need for causal reasoning and showcased how traditional causal methods might be used to test decision-making systems for counterfactual fairness. Similarly, an extensive analysis of counterfactual fairness in several diverse cases by defining different causal models was performed in [20]. Recently, a theoretical study of this notion and its effects from confounding factors was conducted in [22]. The work in [21] provided an extension to this notion of fairness by analyzing more fine-grained *path-specific* effects. The work studied how we can apply counterfactual fairness with respect to *specific paths* in the causal graph, and how we can analyze it. One common point is that all these methods use traditional causal methods for their analyses. Although causal methods are extremely powerful, they also require strict assumptions about the underlying process and any mismatch can lead to high approximation errors. In the situations pertaining to counterfactual fairness, we showed how we can reduce some of the assumptions and use CVAE to approximate the counterfactual generation.

4.2 Deep generative models for causality

There have been several works that have analyzed the intersection of causality and generative models. Such works aim at using generative models to perform causal analysis, or integrate causal reasoning into the model architecture. For instance, the work in [53] proposes a novel integration of causal structure in the GAN model to generate interventional samples for high-dimensional image data. This was extended in [54] to apply causal reasoning for explaining visual models. Related to these works, knowledge of the complete SCM structure has been encoded into VAE models [55], which explores designing a causal latent space in VAEs for image data. Similarly, causal knowledge has also been integrated into VAE and Flow models in [56] for generating interventional image data. The work in [50] works in a more specific situation pertaining to identification of causal effects of certain treatments on the outcome. They utilize VAE models along with TARNets with a simpler causal graph for this purpose. These studies are performed in broader causal scenarios not related to fairness. So, they require more strict knowledge of the causal structures to function.

4.3 Deep generative models in fairness

Recent advancements of generative models have also made them useful tools for conducting fairness studies. Fairness GAN [57] aims to generate synthetic data for multimedia datasets such as images that could help satisfy some notions of fairness for downstream predictions. Another very similar work is FairGAN [58], that also aims to generate data that is fair for some separate predictive system of interest. Fair generative modeling has been studied in [59], where the goal is to generate data that is fair with respect to some demographic information, without any prediction system's outcome in mind. Generative models like GANs have also shown great promise as auditing tools for fairness, e.g., in facial recognition softwares [60].

At the intersection of causality and fairness, there have been some very recent works that explored using generative models for their analysis. Naturally, these works are the most pertinent to the studies conducted in the thesis. Conditional VAE models were used for causal analysis and fairness of treatment in [34]. Methodologically, our work is closest to theirs, however the goal and setups are very different. They study fairness with respect to finding the optimal "treatment" T (medicine) policy with respect to outcome (health) across different sensitive groups. Contrarily, we study *counterfactual fairness* by intervening on A directly to see how observed X and potentially downstream

decisions might change. We consider the prediction of the outcome as separate downstream tasks unlike in [34] which incorporates it into the model. Another work that is closely related is FlipTest [33]. They also work using deep generative models in similar setups as ours without requiring strict causal assumptions. However, there are certain differences. FlipTest tackles the problem of counterfactual mapping by approximating it as task of optimal transport [36]. To solve this, FlipTest uses another class of generative models, generative adversarial networks (GANs), to approximate the mapping of data features between two specific sensitive groups. Thus FlipTest does not estimate the entire data generative process, and has *no latent factor estimation*. This forces FlipTest to train separate models to estimate each individual counterfactual direction. Additionally, FlipTest fails to highlight the important assumptions that need to be satisfied to give meaningful results for fairness. In this thesis, we succinctly showcase the main causal assumptions that have to be satisfied for counterfactual fairness studies. At the same time, as we show, our methodology using CVAEs is more accurate and flexible in practice to study counterfactual fairness.

4.4 Fair representation learning

Our work is also related to those in the space of fair representation learning, a prominent field in fair machine learning that explores preprocessing the data for fairness. One of the seminal works in this space was done in [61]. It explored mapping observed data to "fair" intermediate representations via clustering methods while also providing accurate predictions for a particular task. It solves for the fairness notion of demographic parity [29] without any explicit causal reasoning or analysis for the methodology. Moreover, with unidirectional mapping from data to the latent space, it does not allow for meaningful *interventions* that are required to generate *counterfactuals*. An extension of this work was shown in [62], which uses adversarial learning to purge latent representations of any information related to sensitive features in order to satisfy demographic parity.

Deep generative models have also seen success in this field of fair representation learning. The variational fair autoencoder [63] uses VAEs for fair representation learning without any causal assumptions. They utilize the label Y in the latent representation learning while applying additional Maximum Mean Discrepancy (MMD) regularization for fairness. An attempt to causally interpret the model makes the outcome Y (e.g. success in law school) cause latent factors Z and correspondingly past data features X (e.g. past grades in undergraduate degree). This is an unusual and potentially wrong causal viewpoint. For example, it is very unlikely that success in Law school *after admission*

causes knowledge and past grades. In this work, we model more grounded causal assumptions, because of which we require no additional regularization, while also being able to perform meaningful interventions for fairness.

4.5 Counterfactuals for recourse and explanation

In this thesis, we explore the computation of counterfactuals and using them for the analysis of fairness with respect to decision-making systems. At the same time, there is another parallel space in machine learning research where we can find the use of counterfactuals that is quite different from fairness. This is the domain of algorithmic recourse and explanation of decisions. For a decision-making system that provides a negative outcome to an individual, explanation and recourse allow the individual to understand what he/she would need to change to get a positive outcome. This is also achieved by computing different types of counterfactuals. For instance, the work in [64] provides an alternative approach to generating counterfactual explanations using generative modeling to validate the explanations in the data space. The work in [65] used deep models like VAEs to compute counterfactual explanations that can work for any black-box system that works for tabular data. The recent work in [51] provides a thorough analysis of recourse and counterfactuals from purely a causal perspective. The work provides the need for causal reasoning to guarantee proper recourse, while also showcasing different methods to generate counterfactuals while having different levels of causal knowledge. In this thesis, we do not look into this related field of study, but note it as an important and interesting direction to pursue in the future.

Chapter 5

Experimental Evaluation

In this chapter, we explain the different experiments conducted for a rigorous study of our methods. We perform an extensive analysis to compare how different methods can approximate the counterfactual mappings under the *stated assumptions* (Section 3.1). We compare the deep generative models CVAE and FlipTest that work with fewer assumptions and estimate the SCM forms via function approximation. We compare these models with the causal method shown in [19], the causal MCMC model. These causal methods require strict assumptions about the SCM equations in order to operate. Therefore, we test the MCMC method using various assumptions to understand how mismatches with respect to the true data process impact the approximations. For a thorough study, we use different data setups ranging from synthetic to real-world semi-synthetic, which are described in Section 5.1. We delineate the experimental setups used in Section 5.2. In Section 5.3.1, we use a simple setup to show the relation between causal U and CVAE's estimated Z. The different methods for counterfactual approximation are compared in Section 5.3.2, whereas Section 5.3.3 shows a case study for how these counterfactuals could be used in fairness analysis. Finally, in Section 5.3.4 we highlight how the CVAE model, like the causal model, could be used to train fair downstream predictive models while requiring fewer assumptions.

5.1 Datasets

For thorough experimentation and analysis to compare the different approximation methods, we make use of several synthetic as well as real-world semi-synthetic setups. In the following sections, we explain these setups in more detail.

5.1.1 Synthetic setups

We utilize several synthetic setups to thoroughly compare the approximation performance of the different methods. For each of the synthetic SCMs, we assume 3 observed X, two A which are race (Black, White) and sex (Male, Female). One of the variables, X_3, acts as $X_{\bar{A}}$, not causally affected by A. We consider several different forms of the structural equations for these SCMs to cover multiple scenarios and test which method can be more flexible while providing accurate estimates.

5.1.1.1 Single confounder only

Here, we consider the hidden U in the SCMs comes only through a single confounding source without any additional exogenous variables. This confounder is assumed to come from $U \sim \mathcal{N}(0, 1)$ in all cases. We further explore different forms for the structural equations through various setups.

Linear model Here we consider simple linear additive relationships for each X to U and A.

$$\begin{aligned} X_1 &= b_1 + w_1^U U + w_1^A A \\ X_2 &= b_2 + w_2^U U + w_2^A A \\ X_3 &= b_3 + w_3^U U \end{aligned} \tag{5.1}$$

Non-linear to U We consider that some X are related to U through non-linear functions, but linearly to A.

$$\begin{aligned} X_1 &= \frac{b_1}{c_1 + w_1^e \exp\left(d_1 + w_1^U U\right)} + w_1^A A \\ X_2 &= b_2 + w_2^U \left(1 + U + U^3\right) + w_2^A A \\ X_3 &= b_3 + w_3^U U \end{aligned} \tag{5.2}$$

Non-linear to U, A We consider that some X are related to U as well as A through non-linear functions.

$$\begin{aligned} X_1 &= \frac{b_1}{c_1 + w_1^e \exp\left(d_1 + w_1^U U + w_1^A A\right)} \\ X_2 &= b_2 + w_2^U U + w_2^A A \\ X_3 &= b_3 + w_3^U U \end{aligned} \tag{5.3}$$

5.1.1.2 Single confounder with exogenous variables

For the second setup, we assume that the hidden causal U in the SCM includes, along with some confounder Q, exogenous noise ϵ as well for each X. In all cases, we consider $Q, \epsilon_1, \epsilon_2, \epsilon_3 \sim \mathcal{N}(0, 1)$. As before, we highlight the different forms of the equations we test for here.

Linear model The process is assumed to be linear additive with respect to Q, ϵ and A.

$$\begin{aligned} X_1 &= b_1 + w_1^Q Q + w_1^A A + \sigma_1 \epsilon_1 \\ X_2 &= b_2 + w_2^Q Q + w_2^A A + \sigma_2 \epsilon_2 \\ X_3 &= b_3 + w_3^Q Q + \sigma_3 \epsilon_3 \end{aligned} \tag{5.4}$$

Non-linear to Q We consider that some X are related to Q through non-linear functions, but linearly to A and ϵ.

$$\begin{aligned} X_1 &= \frac{b_1}{c_1 + w_1^e \exp\left(d_1 + w_1^Q Q\right)} + w_1^A A + \sigma_1 \epsilon_1 \\ X_2 &= b_2 + w_2^Q Q + w_2^A A + \sigma_2 \epsilon_2 \\ X_3 &= b_3 + w_3^Q Q + \sigma_3 \epsilon_3 \end{aligned} \tag{5.5}$$

Non-linear, non-additive to U, A Some X are related to Q as well as ϵ through non-linear functions. Also, ϵ is non-additive to A.

$$\begin{aligned} X_1 &= \frac{b_1}{c_1 + w_1^e \exp\left(d_1 + w_1^Q Q + \sigma_1 \epsilon_1\right)} + w_1^A A \\ X_2 &= \left(b_2 + w_2^Q Q + w_2^A A\right)(1 + \sigma_2 \epsilon_2) \\ X_3 &= b_3 + w_3^Q Q + \sigma_3 \epsilon_3 \end{aligned} \tag{5.6}$$

5.1.2 Real-world semi-synthetic setups

We also test the different methods on real-world scenarios. Unfortunately, such data usually lack access to the true generative process. As a result, we cannot directly compare the different methods due to a lack of access to ground-truth counterfactual data. Hence, for these cases, we use a semi-synthetic approach. We use specific SCMs to accurately model the real data and then consider these as the true generative process. This gives us access to real counterfactuals for thorough comparisons.

5.1.2.1 Law school admissions

This setup models the real-world Law school admissions scenario [49]. The causal relations and features are similar to those in [19], detailed in Figure 3.1a and Section 3. The structural equations are listed below:

$$\begin{aligned} \text{LSAT} &= \exp(b_L + w_L^K K + w_L^A A) + \sigma_L \epsilon_L \\ \text{UGPA} &= b_G + w_G^K K + w_G^A A + \sigma_G \epsilon_G \\ \text{FYA} &= w_F^K K + w_F^A A + \epsilon_F \\ K &\sim \mathcal{N}(0,1) \end{aligned} \tag{5.7}$$

Race and sex are the sensitive features A. The confounder K (knowledge of students), along with exogenous noise, affects the observed features: LSAT score and undergraduate GPA (UGPA). The outcome variable is First year average (FYA) the students would have secured *if they were admitted* at the Law school.

5.1.2.2 COMPAS recidivism risk

We consider another semi-synthetic setup for the real-world COMPAS dataset [13]. We use a model similar to that in [20], further described in Section 3 and Figure 3.1b. The structural equations are shown here that model the real-world data.

$$\begin{aligned} J_F &= b_{J_F} + w_{J_F}^J U_J + w_{J_F}^E E + w_{J_F}^A A + \sigma_{J_F} \epsilon_{J_F} \\ J_M &= b_{J_M} + w_{J_M}^J U_J + w_{J_M}^E E + w_{J_M}^A A + \sigma_{J_F} \epsilon_{J_M} \\ P &= b_P + w_P^D U_D + w_P^E E + w_P^A A + \sigma_P \epsilon_P \\ C &= \text{sigmoid}(b_C + w_C^D U_D + w_C^E E + w_C^A A) \\ R &= b_R + w_R^D U_D + w_R^E E + w_R^A A + \sigma_R \epsilon_R \\ U_D &\sim \mathcal{N}(0,1);\ U_J \sim \mathcal{N}(0,1) \end{aligned} \tag{5.8}$$

The sensitive attribute A is race (Black, White). Age E along with A affect all features: Crime (C), Priors (P), Juvenile felony J_F, Juvenile misdemeanour J_M. There are two confounders U_D, U_J that indicate criminality, along with exogenous variables. R represents the potential risk of criminal defendants to recidivate. It is the target variable for any downstream predictive system.

5.2 Experimental Setup

Deep model setups Both CVAE and FlipTest are modeled using deep neural networks with 3 hidden layers and 128 neurons per layer. We tuned the hyperparameters (σ^2 for CVAE, λ for FlipTest) to select the best models. For CVAE, the latent dimension of Z was set the same as the number of observed features to model all noise factors. As the CVAE model gives a probabilistic estimate of Z for each individual, we compare against the true deterministic counterfactuals by taking the mean outcome over 500 Monte-Carlo samples of Z.

Causal MCMC setups We further test causal methods that employ MCMC [19]. These require additional knowledge about the structural equations' forms. So, we train these models with *varying assumptions* about the equations. These help to understand the effects on the causal models when these assumptions are misaligned with respect to the true process. *Oracle* models $\mathcal{M}_*$ have knowledge of the true structural equations' forms f_*, be it linear or non-linear. In such cases, it knows that $X_A = f_*(U, A)$ and $X_{\bar{A}} = f_*(U)$. Given this knowledge of the exact form, the model can learn the parameters and the hidden factors from the data. At the same time, we consider some setups where the MCMC models make wrong assumptions about the data process. For nonlinear synthetic setups, causal models making the wrong assumption of linearity are considered. They are denoted by $\mathcal{M}_{\text{Lin}}$, where it considers $X_A = b_{X_A} + w^U_{X_A} U + w^A_{X_A} A$ and $X_{\bar{A}} = b_{X_{\bar{A}}} + w^U_{X_{\bar{A}}} U$. Furthermore, in our synthetic setup, not all X are causally affected by A. We test $\mathcal{M}^A$ that make the wrong assumption of all X being causally dependent on A, or, $X = f(U, A); \forall X \in \{X_A, X_{\bar{A}}\}$. All these MCMC models are trained using probabilistic modeling in Stan [66].

Downstream model setups Downstream predictive systems in Sections 5.3.3 and 5.3.4 are trained as linear regression models with L_2 weight penalization. These models are trained using the scikit-learn package with default solvers.

5.3 Experiments

Now, we explain and discuss the main evaluations we perform. These are aimed at analyzing our hypothesis regarding counterfactual approximations for fairness with fewer assumptions. We also highlight one potential application of such counterfactuals for analyzing fairness: auditing for counterfactual fairness. Finally, we explore how we can use these approximation methods to train fair predictors.

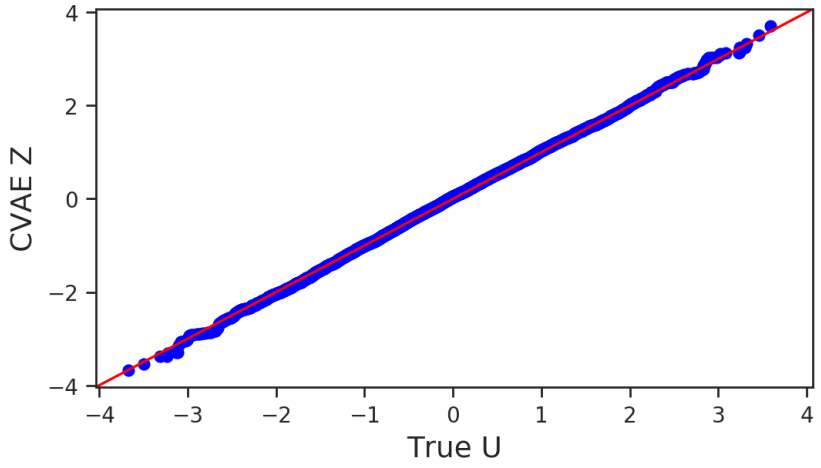


Figure 5.1: Comparing true confounder U with estimates through CVAE latent Z for simple synthetic setup (Eqn. 5.1) using QQ-plot. We can see that Z manages to capture the true U faithfully.

5.3.1 What does the latent Z in CVAE learn?

As discussed in Section 3.2, we approximate the data generative process using CVAE via a latent distribution space Z. However, *how does the latent Z of the CVAE model relate to the hidden exogenous U of the causal process?* To answer this, we conduct a simple analysis with a single noise source using the synthetic setup of Equation 5.1. We fit the CVAE on this data, then compare the latent Z to the true U of the underlying causal process. Figure 5.1 shows how the CVAE is able to model hidden noise factors of a causal process through Z, incurring very little estimation error. The mean estimation error between U and Z is **0.013**. For perspective, oracle $\mathcal{M}_*$ makes an error of **0.01**. This clearly highlights how the CVAE can be employed in causal setups, and can learn causal effects for counterfactual fairness studies. We hypothesize that for multiple noise sources, the CVAE Z would learn an entangled representation of the same. The CVAE model would still be able to estimate the *total* effect of the causal U on the data generation process. As shown next, under the *stated assumptions*, this still allows CVAE to make good estimates for counterfactual mappings that are useful for the analysis of *counterfactual fairness*.

5.3.2 Approximating causal counterfactuals

We analyze how well the different methods approximate counterfactuals for fairness scenarios postulated in Section 3.1. Access to the ground-truth SCMs in Section 5.1 allows us to generate *true* counterfactuals to compare against. For data X and counterfactual

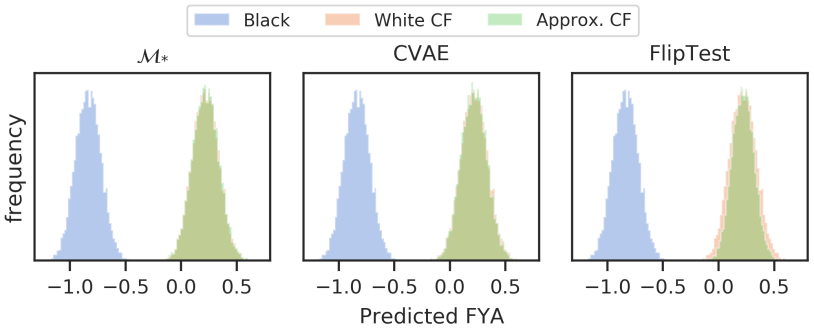
Table 5.1: Counterfactual quantity approximation in different synthetic setups for Black to White transformation. CVAE manages to approximate reasonably well compared to the *ideal* models $\mathcal{M}_*$, beating out FlipTest in all cases. However, wrong assumptions in MCMC models lead to considerably higher errors.

Experiment Setup	**Model**	**Mean Error**
Linear (conf. only)	$\mathcal{M}_*$	0.002 ± 0.0001
	$\mathcal{M}_*^A$	0.007 ± 0.002
	FlipTest	0.007 ± 0.001
	CVAE	0.006 ± 0.002
Non-linear to U (conf. only)	$\mathcal{M}_*$	0.006 ± 0.004
	$\mathcal{M}_*^A$	0.008 ± 0.006
	$\mathcal{M}_{\text{Lin}}$	0.067 ± 0.001
	$\mathcal{M}_{\text{Lin}}^A$	0.067 ± 0.0001
	FlipTest	0.034 ± 0.009
	CVAE	0.008 ± 0.001
Non-linear to U, A (conf. only)	$\mathcal{M}_*$	0.012 ± 0.01
	$\mathcal{M}_*^A$	0.02 ± 0.013
	$\mathcal{M}_{\text{Lin}}$	0.136 ± 0.0002
	$\mathcal{M}_{\text{Lin}}^A$	0.136 ± 0.001
	FlipTest	0.068 ± 0.013
	CVAE	0.025 ± 0.01
Linear (conf. + exog.)	$\mathcal{M}_*$	0.004 ± 0.003
	$\mathcal{M}_*^A$	0.05 ± 0.032
	FlipTest	0.013 ± 0.005
	CVAE	0.006 ± 0.0016
Non-linear to U (conf. + exog.)	$\mathcal{M}_*$	0.0035 ± 0.0005
	$\mathcal{M}_*^A$	0.005 ± 0.001
	$\mathcal{M}_{\text{Lin}}$	0.035 ± 0.012
	$\mathcal{M}_{\text{Lin}}^A$	0.042 ± 0.009
	FlipTest	0.033 ± 0.007
	CVAE	0.008 ± 0.002
Non-linear, Non-additive to U, A (conf. + exog.)	$\mathcal{M}_*$	0.022 ± 0.002
	$\mathcal{M}_*^A$	0.029 ± 0.009
	$\mathcal{M}_{\text{Lin}}$	0.023 ± 0.005
	$\mathcal{M}_{\text{Lin}}^A$	0.03 ± 0.01
	FlipTest	0.042 ± 0.004
	CVAE	0.021 ± 0.001
Law school admissions	$\mathcal{M}_*$	0.27 ± 0.001
	$\mathcal{M}_{\text{Lin}}$	0.32 ± 0.02
	FlipTest	0.3 ± 0.018
	CVAE	0.25 ± 0.011
COMPAS recidivism risk	$\mathcal{M}_*$	0.035 ± 0.018
	$\mathcal{M}_{\text{Lin}}$	0.17 ± 0.05
	FlipTest	0.12 ± 0.016
	CVAE	0.06 ± 0.012

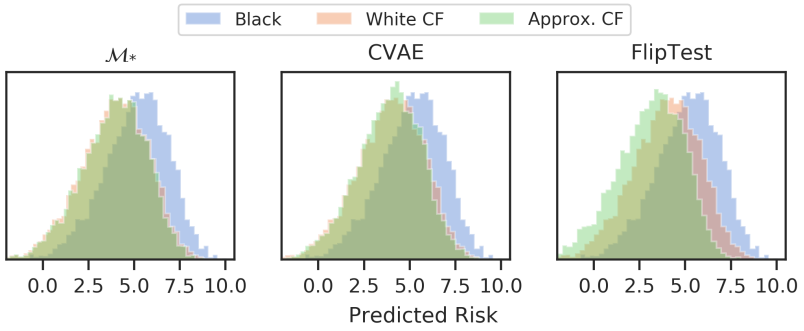
X^c, we compute estimation error Err $= \frac{1}{N}\sum_{i=1}^{N}|X_i - X_i^c|$. The results are surmised in Table 5.1. The *oracle* causal $\mathcal{M}_*$ naturally reports the lowest errors as it has the most knowledge of the causal processes. But we also show how sensitive these causal models are to the assumptions they work with. For example, the misspecified models $\mathcal{M}_*^A$ incur 30% more error, whereas $\mathcal{M}_{\text{Lin}}$ report almost 10 times higher error than the oracle models. Causal models $\mathcal{M}_{\text{Lin}}^A$ that make both wrong assumptions on linearity and for A end up being the worst performers in approximating the true counterfactuals. In comparison to these causal methods, the deep models make fewer assumptions about the SCM. However, the FlipTest models usually incur high errors, in many cases, being significantly worse than the oracle models. This might be because FlipTest fails to capture the entire data-generative process and the total cause of the latent factors. It only models the transformation of features between two groups. At the same time, CVAE manages to come quite close to the performance of the oracle $\mathcal{M}_*$ in all the experimental setups. Hence, for *counterfactual fairness*, we show how the CVAE is able to generate faithful counterfactuals despite working with less causal knowledge.

5.3.3 Case Study: Fairness auditing with counterfactuals

In the previous analysis, we compared the different methods for counterfactual approximation in a variety of synthetic and real-world semi-synthetic setups. However, *how can these counterfactuals be used for practical fairness analysis?* In this section, we highlight one such potential application, and compare the performance of the different counterfactual approximation methods in the specific use case. Such generated counterfactuals can actually be used to audit downstream predictive models for *counterfactual fairness*. To highlight this, we consider two semi-synthetic setups that model real-world data for (a) Law school admissions and (b) COMPAS recidivism risk prediction. We train downstream regression models that predict (a) the likely first year average (FYA) for Law school applicants and (b) possible recidivism risk of criminal defendants. Both of these models utilize all observed features of the data to provide predictions. Now, we aim to audit these models to see if they conform to the notion of counterfactual fairness, while at the same time comparing the different counterfactual approximation methods for this auditing task. To audit them, oracle $\mathcal{M}_*$, CVAE and FlipTest are used to generate approximate White counterfactuals for Black individuals. But, *how do we know which method can be trusted more for auditing?* For this, we compare them against auditing with true counterfactuals generated directly using Equations 5.7 and 5.8. In Figure 5.2, we see that all the methods highlight potential bias in both predictive models with respect to race. However, as per the true auditing results, FlipTest gives a different report, especially for COMPAS (Figure 5.2b), where it reports higher bias than



(a) Law school admissions



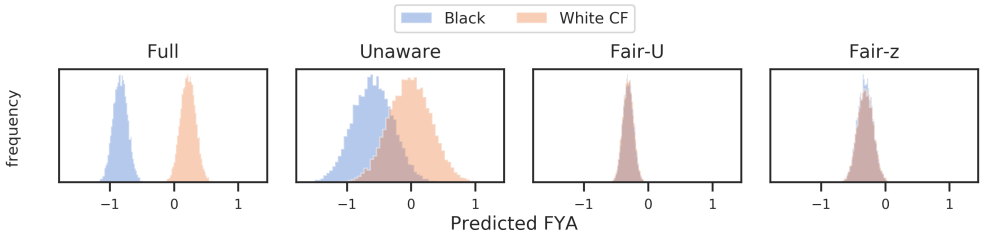
(b) COMPAS recidivism risk

Figure 5.2: Auditing predictive systems for (a) FYA in Law school admissions and (b) recidivism risk of criminal defendants in COMPAS. Semi-synthetic analysis performed with respect to Black-White counterfactual mapping. Each approximation method compared with ground-truth counterfactuals (CFs) from Equations 5.7 and 5.8. CVAE, similar to $\mathcal{M}_*$, matches ground-truth auditing results and is more trustworthy. But FlipTest diverges from true auditing results, particularly in (b), reporting more bias in audited model than what truly exists.

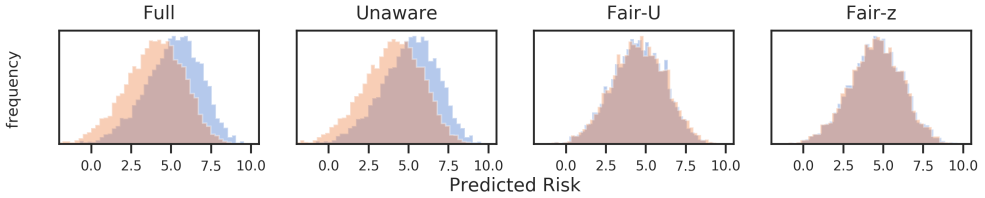
is actually present in the predictive model. Despite working with less strict assumptions, the CVAE matches the causal oracle $\mathcal{M}_*$ in being faithful to the ground-truth auditing results. In conclusion, generative models like the CVAE can be trustworthy auditing tools for fairness, performing as well as oracle causal methods while working with significantly fewer causal assumptions.

5.3.4 Training counterfactually fair predictors

Finally, we explore how the different counterfactual approximation methods could be used directly to train downstream counterfactually fair predictors. We again use the semi-synthetic setups as in Section 5.3.3. The task is to *fairly predict* the target variable, in these cases, (a) FYA for Law school and (b) recidivism risk for COMPAS, such that the prediction for any data point and its counterfactual are the same. For both cases, we utilize black individuals and their white counterfactuals to test fairness. For a rigorous



(a) Law school admissions



(b) COMPAS recidivism risk

Figure 5.3: Training counterfactually fair predictor for (a) FYA of Law school and (b) recidivism risk for COMPAS. *Fair-U* uses the latent factors learnt by the oracle causal $\mathcal{M}_*$. *Fair-z* uses latent representations of CVAE . Baseline models are *Full* (use all observed features), *Unaware* (use all observed features except A). Models audited using true causal counterfactuals using Equations 5.7 and 5.8. Visualizing predictions for Black individuals and true White counterfactuals. *Fair-z* model as well as ideal *Fair-U* provide fair predictions with respect to the counterfactuals.

comparison, we use the true counterfactuals computed directly from Equations 5.7 and 5.8. Because of our model choices for the CVAE, we can use the latent Z that is independent of the sensitive feature information to directly train a fair predictive model *Fair-z*. We use the hidden causal U learnt by the oracle $\mathcal{M}_*$ to train the model *Fair-U*, which are the *Level-2* models in [19]. Owing to a lack of modeling latent factors of the data, we cannot use FlipTest directly to train fair models. As baseline systems, we consider a *Full* model that uses all X, A, and an *Unaware* model that uses only X. We measure the prediction error each model incurs using root mean squared error (RMSE). The mean prediction difference between any individual ($\hat{Y}$) and its counterfactual ($\hat{Y}^c$) is measured as unfairness $= \frac{1}{N}\sum_{i=1}^{N}\left|\hat{Y}_i - \hat{Y}_i^c\right|$.

Figure 5.3 shows how *Full* and *Unaware* models are biased in its predictions. In comparison *Fair-z* manages to provide almost same predictions for both black and their causal white counterfactuals, performing at par with the ideal *Fair-U*. Table 5.2 provides a more fine-grained analysis and shows the accuracy-fairness tradeoff with respect to the *biased* ground-truth labels. As expected, *Full* and *Unaware* models are highly accurate while being more biased. On the other hand, *Fair-z* and *Fair-U* perform similarly: they incur more predictive error but are significantly more fair with respect to the counterfactuals. Through this experiment, we show how the CVAE manages to learn counterfactually fair representations. This allows using it to train downstream predictive models that,

Table 5.2: Root mean squared error (RMSE) and unfairness w.r.t. counterfactuals in different models trained on downstream predictive tasks. Comparing models Full (use X, A), Unaware (use X), Fair-U (causal U from $\mathcal{M}_*$), Fair-z (latent Z of CVAE). The latter two models achieve low unfairness while incurring some additional error against biased ground-truth labels.

Dataset	**Model**	**RMSE**	**Unfairness**
Law school admissions	Full	1	1.05
	Unaware	1.04	0.58
	Fair-U	1.12	0.01
	Fair-z	1.12	0.01
COMPAS recidivism risk	Full	1.86	1.2
	Unaware	1.86	1.1
	Fair-U	2.19	0.05
	Fair-z	2.21	0.05

like the ideal causal *Fair-U*, satisfy the notion of *counterfactual fairness*, while relying on markedly less causal knowledge.

5.4 Discussion

Through our extensive evaluations, we highlight that for counterfactual fairness, we do not require traditional causal methods and their strong assumptions to approximate counterfactual quantities. Through a series of rigorous synthetic and real-world semi-synthetic experiments, we show that traditional causal methods are indeed prone to making significant approximation errors when their strong assumptions do not align with the true data process. Indeed, the main assumptions listed in Section 3.1 are sufficient for faithful counterfactual computation. We go on to confirm our hypothesis that we *can* use deep latent variable generative models such as CVAEs for the purpose. We show that CVAEs can approximate the total effect of hidden causal factors. Additionally, CVAEs can also generate faithful counterfactuals while requiring fewer assumptions than the oracle causal methods. Through a case study of auditing for counterfactual fairness, we show how counterfactuals could be used for real-world fairness analysis. We demonstrate that not all approximation methods can give robust auditing reports; it is important to model the entire data process faithfully. Likewise, it is not required to make strong assumptions, as the CVAE gives robust auditing results matching the oracle causal method requiring perfect knowledge of the true process. Finally, we show that strict causal assumptions and traditional causal methods are not essential to train counterfactually fair predictors. CVAE models working with *only* the main assumptions can also be used for the purpose, providing the same performance as the causal methods.

Chapter 6

Conclusion and Future work

6.1 Conclusion

Causal analysis in the study of algorithmic fairness is growing to become a very important avenue of research. Recent works have led to various notions of causal fairness, the first and most popular of which is the notion of counterfactual fairness. Evaluating this notion of fairness for algorithmic decision-making systems requires the generation of faithful counterfactual quantities for actual data-points. Traditionally, this has been studied using causal methods. Unfortunately, these methods require strict assumptions about the underlying data process, including both the causal graph and the structural equations corresponding to *all* the features. However, this knowledge may not be available in most practical settings, particularly in societal settings where we wish to study fairness. As we have shown, misalignment with respect to these considerations often results in the causal methods making significant errors in the counterfactual approximation process. This makes the study of counterfactual fairness practically challenging.

In this thesis, we demonstrate that for analyzing counterfactual fairness, we do not need to work with such strict assumptions. We concisely highlight the main considerations that are made about the data for most fairness analysis. We go over the extensive literature related to counterfactual fairness and confirm that these assumptions are valid and implicitly encompass all these pertinent scenarios. We remarked how this ends up allowing us to reduce the supplementary causal assumptions. Indeed, for counterfactual fairness, we only need to consider the highlighted main assumptions. Thus, we do not actually require traditional causal methods and their strict assumptions to study this notion of fairness. We show how deep generative models, in particular, latent variable models like CVAEs, can instead be used in practice with these reduced considerations. Through rigorous evaluation, we show CVAEs can estimate the total effect from causal hidden factors and faithfully approximate the counterfactual process for fairness, matching the performance of the ideal causal methods. Through a simple

case study for auditing fairness, we demonstrate that the CVAE acts as a more reliable model for analyzing counterfactual fairness, despite working with fewer assumptions. These results also highlighted that not all generative models can work as well in these settings. In fact, as we see, FlipTest failed to approximate the counterfactual process effectively that also led it to give less robust auditing reports. Finally, we show how the CVAE, like an oracle causal model, can also be used to train a counterfactually fair predictive model, while requiring less causal information regarding the data process. In conclusion, we showed that it *is* possible to practically operationalize counterfactual fairness. We can reduce the significant causal assumptions made by traditional methods to only a few main considerations. This allows using deep generative models like CVAE to faithfully approximate counterfactuals and even directly train predictive models that can satisfy the notion of counterfactual fairness.

6.2 Future Work

Although the results show promise, there are several open avenues to explore in this space. We delineated the specific assumptions that are required for our analysis and explained how they were prevalent in fairness studies. However, there can be specific circumstances where the assumptions may not hold. It would be interesting to study the fairness implications of the different methods if these assumptions do not hold. As mentioned earlier, it can happen that there are certain confounding factors in the generative process that are not well approximable from the observed features. Clearly, our method would not be able to handle such situations, and the fairness consequences of these would require more careful analysis. Traditional fairness and causal studies also do not assume hidden confounding between the sensitive features and the hidden causal factors. However, there might be certain scenarios where such confounding would exist and be considered for fairness analyses. Clearly, this would violate the monotonicity of features across the sensitive groups that we explain in Chapter 3. Failing to handle this might lead to unfair results in these circumstances. As a result, such cases would need detailed formulation and more rigorous studies. In this thesis, we explore studying the practical analysis of counterfactual fairness. Recently, there have been multiple extensions and related notions introduced in the space of causality and fairness [21, 30]. Working with these notions would require further causal knowledge to be considered. Encoding this knowledge into deep generative models, e.g., distinguishing between descendants and nondescendants of sensitive features to investigate path-specific effects and specifically considering resolving variables would make for a fascinating analysis. Causal methods have also traditionally been studied from a purely theoretical perspective. The analysis regarding the human interpretation of such mappings have been mostly ignored thus

far. Taking inspiration from recent works [67, 68], it would be interesting to study how humans or experts rate the counterfactual mappings generated by these methods for the purpose of counterfactual fairness. Such an analysis could provide vital insights and go a long way to allowing real-world deployment of causal analysis in the fairness of decision-making systems. Related to this, it would also be interesting to explore the possibility of designing a human-in-the-loop model for providing a causal analysis of fairness. At the same time, using sensitive features and social constructs in causal models and actually performing interventions on the same needs further discussion across different communities. We provided a brief discussion on this topic, but questions such as the inclusion of race or gender in these analyses, proper operationalization of these attributes in the causal model require further insight. Finally, the actual deployment of causal methods for fairness also needs to be studied from the legal perspectives of bias, discrimination and fairness. As we briefly discussed, there have been some preliminary exploration and discussion in this space. But, we believe there is a great scope in interdisciplinary research to analyze and operationalize such causal methods to comply with legal notions of antidiscrimination. We believe this work opens new and interesting research directions to pursue in causal fairness for real-world scenarios.

List of Figures

List of Tables

Bibliography

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[2] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[3] Smart Compose: Using Neural Networks to Help Write Emails, 2018. `https://ai.googleblog.com/2018/05/smart-compose-using-neural-networks-to.html`.

[4] Sahin Cem Geyik, Vijay Dialani, Meng Meng, and Ryan Smith. In-session personalization for talent search. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2107–2115, 2018.

[5] Viet Ha-Thuc, Ye Xu, Satya Pradeep Kanduri, Xianren Wu, Vijay Dialani, Yan Yan, Abhishek Gupta, and Shakti Sinha. Search by ideal candidates: Next generation of talent search at linkedin. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 195–198, 2016.

[6] Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40, 2009.

[7] Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

[8] Rohan Bhardwaj, Ankita R Nambiar, and Debojyoti Dutta. A study of machine learning in healthcare. In *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 236–241. IEEE, 2017.

[9] Kate Crawford and Ryan Calo. There is a blind spot in ai research. *Nature*, 538 (7625):311–313, 2016.

[10] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[11] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China, 2019. Association for Computational Linguistics.

[12] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578, 2017.

[13] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*, 23, 2016.

[14] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[15] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[16] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[18] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[19] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in neural information processing systems*, pages 4066–4076, 2017.

[20] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in neural information processing systems*, pages 6414–6423, 2017.

[21] Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

[22] 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.

[23] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

[24] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[25] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491, 2015.

[26] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision*, pages 835–851. Springer, 2016.

[27] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2, 2016.

[28] Wiji Arulampalam, Alison L Booth, and Mark L Bryan. Is there a glass ceiling over europe? exploring the gender pay gap across the wage distribution. *ILR Review*, 60(2):163–186, 2007.

[29] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.

[30] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[31] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal

reasoning. In *Advances in neural information processing systems*, pages 656–666, 2017.

[32] Don Van Ravenzwaaij, Pete Cassey, and Scott D Brown. A simple introduction to markov chain monte–carlo sampling. *Psychonomic bulletin & review*, 25(1):143–154, 2018.

[33] Emily Black, Samuel Yeom, and Matt Fredrikson. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121, 2020.

[34] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 349–358, 2019.

[35] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

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

[37] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[38] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[39] Issa Kohler-Hausmann. Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, 113:1163, 2018.

[40] Maya Sen and Omar Wasow. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19, 2016.

[41] Lily Hu and Issa Kohler-Hausmann. What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 513–513, 2020.

[42] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[43] Clark Glymour. Comment: Statistics and metaphysics. *Journal of the American Statistical Association*, 81(396):964–966, 1986.

[44] Donald B Rubin. Comment: Which ifs have causal answers. *Journal of the American statistical association*, 81(396):961–962, 1986.

[45] Margaret Mooney Marini and Burton Singer. Causality in the social sciences. *Sociological methodology*, 18:347–409, 1988.

[46] Robert N Strassfeld. If...: Counterfactuals in the law. *Geo. Wash. L. Rev.*, 60:339, 1991.

[47] Sheila R Foster. Causation in antidiscrimination law: Beyond intent versus impact. *Hous. L. Rev.*, 41:1469, 2004.

[48] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[49] Linda F Wightman and Henry Ramsey. *LSAC national longitudinal bar passage study*. Law School Admission Council, 1998.

[50] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in neural information processing systems*, pages 6446–6456, 2017.

[51] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. *arXiv preprint arXiv:2002.06278*, 2020.

[52] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR (Poster)*. OpenReview.net, 2017.

[53] Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018.

[54] Á. Parafita and J. Vitrià. Explaining visual models by causal attribution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4167–4175, 2019. doi: 10.1109/ICCVW.2019.00512.

[55] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. *arXiv preprint arXiv:2004.08697*, 2020.

[56] Nick Pawlowski, Daniel C Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *arXiv preprint arXiv:2006.06485*, 2020.

[57] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3:1–3:9, 2019. doi: 10.1147/JRD.2019.2945519.

[58] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE, 2018.

[59] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1887–1898. PMLR, 13–18 Jul 2020.

[60] Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439*, 2019.

[61] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[62] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[63] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[64] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

[65] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pages 3126–3132, 2020.

[66] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[67] David Madras, Toniann Pitassi, and Richard S. Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *NeurIPS*, pages 6150–6160, 2018.

[68] Yahav Bechavod, Christopher Jung, and Steven Z. Wu. Metric-free individual fairness in online learning. In *NeurIPS*, 2020.

Appendix A

Dataset details

The datasets used in Chapter 5 include both synthetic as well as real-world semisynthetic data. Working with these datasets helps us conduct a thorough analysis for comparing our method and the other baseline methods for the task of counterfactual approximation. We delineate the different details for each of the datasets used in the evaluations of Chapter 5. These details highlight the variety in the different datasets we have used to show that CVAE models can be flexible in generating faithful counterfactuals. For each dataset, we report some important statistics for each of the features like the mean, standard deviation, min-max values, quantile values, etc. Furthermore, through detailed pair plots, we visualize the relations between each pair of features in the data. These are aimed at providing a deeper insight into the different datasets used and the related complexities present in each.

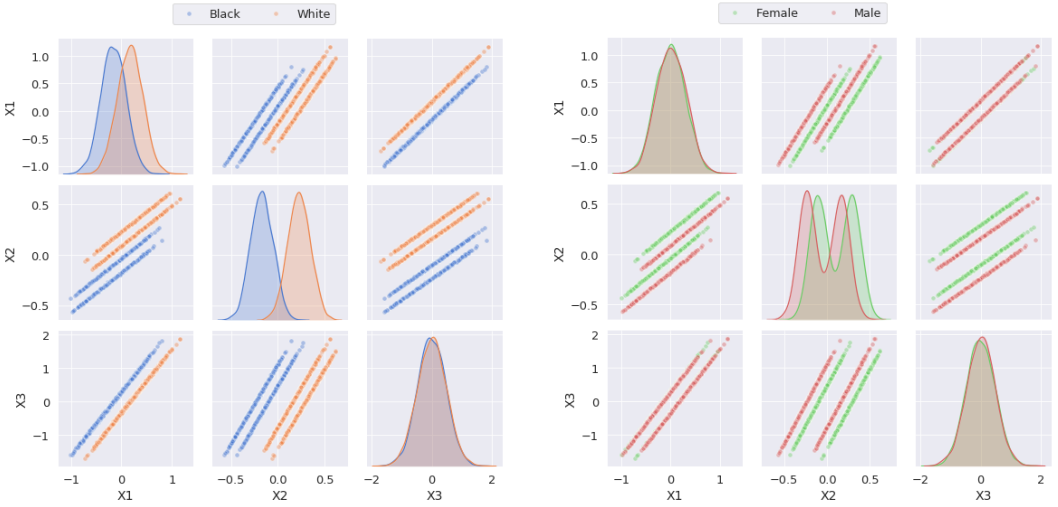
A.1 Synthetic datasets with only single confounder

Linear relationships with confounder and sensitive features This dataset is generated with additive linear relationship between the sensitive A, confounder U and the features X following the process in Equation 5.1. The details and distributions are shown in Table A.1 and Figure A.1.

Non-linear relationships with confounder This dataset is generated such that the structural equations involve nonlinearity with respect to the confounder. The form of the relationships used is given in Equation 5.2. We consider exponential as well as polynomial relations in the equations. The distribution and statistics are shown in Table A.2 and Figure A.2.

Table A.1: Feature statistics for the synthetic dataset with linearity relations and single confounder

	X1	X2	X3
mean	0.004644	0.029152	-0.000663
std	0.299870	0.233548	0.471979
min	-1.181656	-0.647792	-1.956550
25%	-0.205945	-0.171136	-0.318425
50%	0.005846	0.025852	0.000696
75%	0.214951	0.230226	0.322070
max	1.220672	0.722659	2.017464



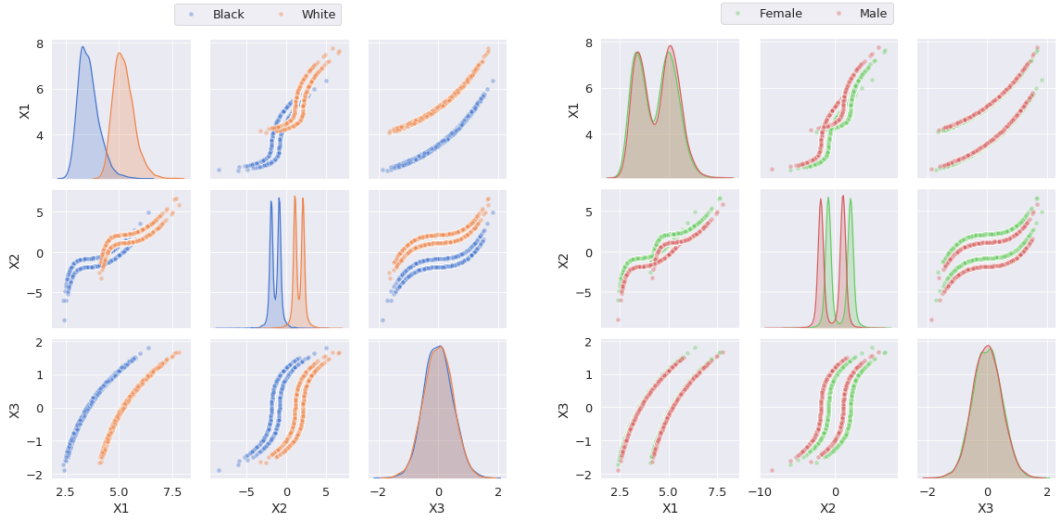
(a) Respect to race

(b) Respect to sex

Figure A.1: Density and scatter plots for pairwise feature relationships in the synthetic setup with single confounder and linear relationships

Table A.2: Feature statistics for the synthetic dataset with single confounder which causes the features through a nonlinear relationship

	X1	X2	X3
mean	4.403591	0.149431	0.001136
std	0.949136	1.643546	0.474402
min	2.282333	-9.567023	-2.064378
25%	3.529664	-1.360167	-0.319717
50%	4.512809	0.059084	-0.000583
75%	5.171474	1.640677	0.322390
max	8.121509	8.963523	2.222603



(a) Respect to race (b) Respect to sex

Figure A.2: Density and scatter plots for pairwise feature relationships in the synthetic setup with single confounder and nonlinear relationship involving it

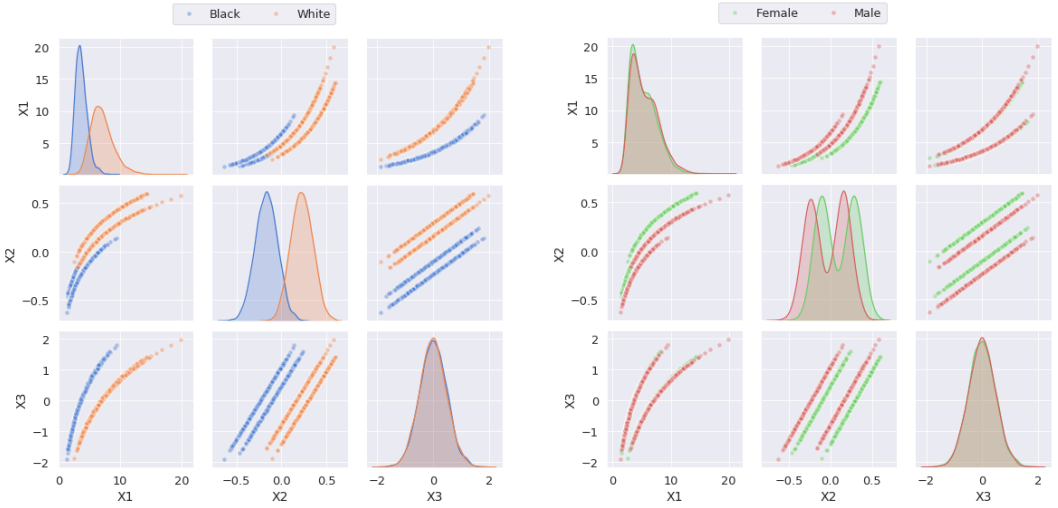
Table A.3: Feature statistics for the synthetic dataset with single confounder with nonlinear relationships involving the confounder and sensitive features

	X1	X2	X3
mean	5.426447	0.030369	0.001370
std	2.238232	0.231782	0.479752
min	1.353508	-0.633118	-2.003764
25%	3.595417	-0.167122	-0.319703
50%	5.005536	0.030252	0.000487
75%	6.930331	0.228641	0.323620
max	19.926446	0.637505	2.277124

Non-linear relationships with confounder and sensitive features This dataset is generated with structural equations that involve nonlinearity with respect to the confounder as well as sensitive features. We consider exponential relations with respect to the confounder and sensitive attributes. The form of the equations used is given in Equation 5.3. The feature statistics and inter feature distributions are shown in Table A.3 and Figure A.3.

A.2 Synthetic datasets with confounder and exogenous noise

Linear relationships with hidden factors and sensitive features This dataset is generated with additive linear relationship between the sensitive A, hidden factors U (involving a confounder and exogenous noise variables) and the features X following the



(a) Respect to race (b) Respect to sex

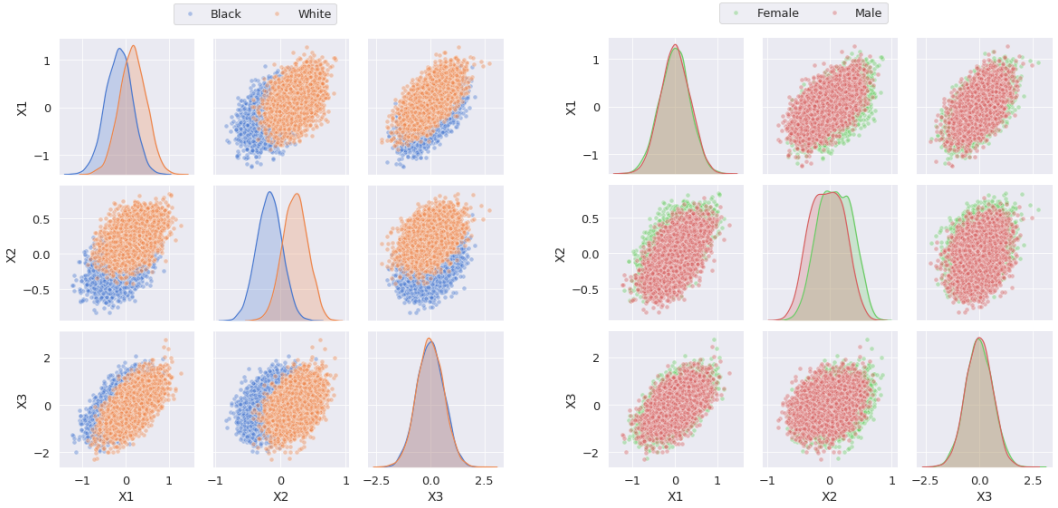
Figure A.3: Density and scatter plots for pairwise feature relationships in the synthetic setup with single confounder and nonlinear relationship involving it and sensitive features

Table A.4: Feature statistics for the synthetic dataset with single confounder and exogenous noise with linear relationships with respect to them.

	X1	X2	X3
mean	0.003443	0.028149	-0.002061
std	0.347067	0.275881	0.617425
min	-1.291171	-0.852258	-2.481044
25%	-0.232374	-0.177197	-0.425528
50%	0.004794	0.027630	-0.002159
75%	0.238804	0.236295	0.417868
max	1.269363	0.965580	2.744171

process in Equation 5.4. The feature statistics are shown in Table A.4 while the inter feature characteristics are demonstrated in Figure A.4.

Non-linear relationships with confounder This dataset is generated such that the structural equations involve nonlinearity with respect to the confounding variable in the causal process. The form of the relationships used is given in Equation 5.5. We consider exponential relations in the equations. The feature characteristics can be seen in Table A.5 and Figure A.5.



(a) Respect to race (b) Respect to sex

Figure A.4: Density and scatter plots for pairwise feature relationships in the synthetic setup with linear relationships regarding the hidden factors (confounder and exogenous noise) and sensitive features

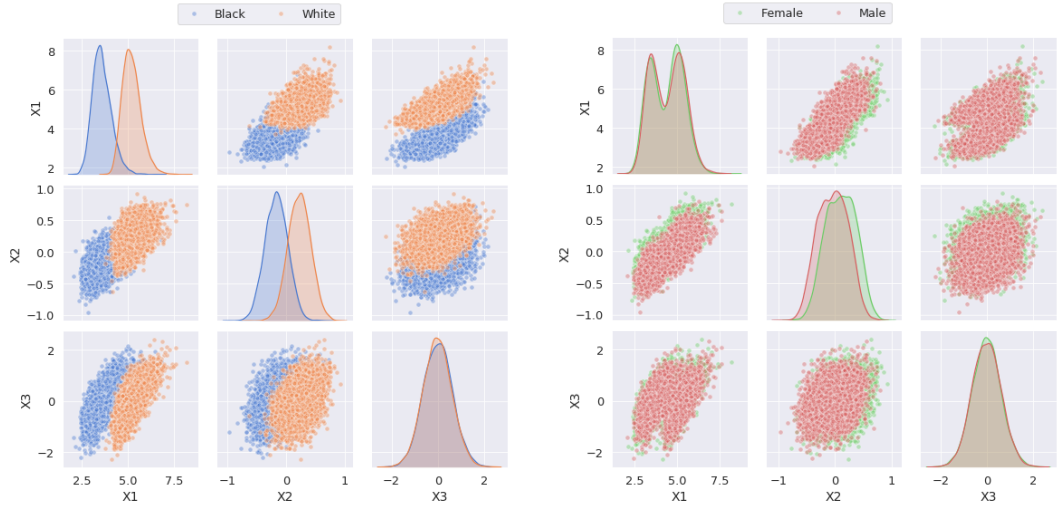
Table A.5: Feature statistics for the synthetic dataset with confounder and exogenous noise. The confounder affects the data through non-linear relation.

	X1	X2	X3
mean	4.409812	0.031795	-0.002476
std	0.964339	0.275249	0.617451
min	2.068358	-0.967109	-2.397208
25%	3.541113	-0.175395	-0.416172
50%	4.498690	0.032719	-0.003936
75%	5.186337	0.238650	0.410765
max	8.196933	0.929618	2.373112

Non-linear relationships with hidden factors and sensitive features This dataset is generated with structural equations that involve nonlinearity with respect to the hidden factors (confounder and exogenous variables) as well as sensitive features. We consider exponential relations. We also consider non-linear relations between the hidden factors and the sensitive features for added complexity. The form of the equations used is given in Equation 5.6. Feature statistics and characteristics can be seen in Table A.6 and Figure A.6.

A.3 Real-world semisynthetic datasets

Law school admission This dataset includes the test scores of candidates who are applying to law school after their undergraduate education. The sensitive features are



(a) Respect to race

(b) Respect to sex

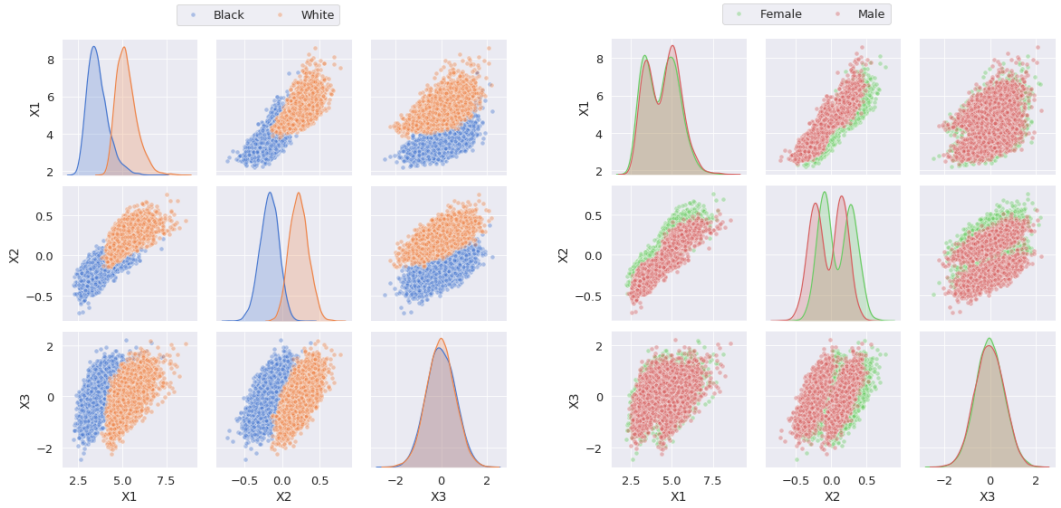
Figure A.5: Density and scatter plots for pairwise feature relationships in the synthetic setup with confounder and exogenous noise. The confounder causes features through nonlinear functions

Table A.6: Feature statistics for the synthetic dataset with confounder and exogenous noise. There exist non-linear relations to the hidden factors and sensitive features

	X1	X2	X3
mean	4.434719	0.030075	-0.002946
std	1.017618	0.233784	0.618176
min	2.131410	-0.722513	-2.727280
25%	3.525159	-0.163896	-0.420502
50%	4.516270	0.028678	-0.001447
75%	5.194929	0.223937	0.412841
max	9.333748	0.833448	2.499900

race (Black-White) and sex (Male-Female). The other features are LSAT score, and undergraduate GPA (UGPA). The outcome variable is the first year average (FYA) that admitted students achieve in law school. Refer to Table A.7 and Figure A.7 for a detailed report of the feature statistics and visualizations.

COMPAS recidivism risk prediction This dataset includes the data of criminal inmates and their potential risk for recidivism. In this dataset, race (White-Black) is considered to be the sensitive feature with respect to which we need to achieve fairness. The different observed features are juvenile felony (JuvFel), juvenile misdemeanor (JuvMisd), Crime, Priors, Age. The outcome variable is the risk of recidivism that each inmate poses, that is to be predicted from the rest of the features. The statistics for each of the features and feature relationship visualizations can be seen in Table A.8 and Figure A.8 respectively.

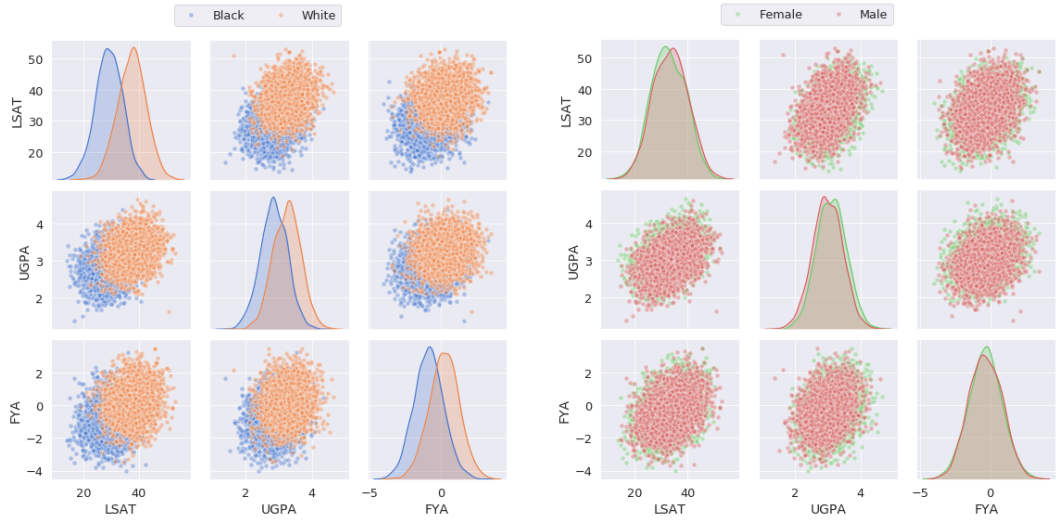


(a) Respect to race (b) Respect to sex

Figure A.6: Density and scatter plots for pairwise feature relationships in the synthetic setup with confounder and exogenous variables where nonlinear relationships exist with respect to them and sensitive features

Table A.7: Feature statistics for the Law school admission dataset

	LSAT	UGPA	FYA
mean	33.323431	3.069783	-0.313463
std	6.453742	0.450226	1.153091
min	10.603190	1.375222	-4.830155
25%	28.659733	2.763090	-1.102393
50%	33.228275	3.070528	-0.312264
75%	37.929004	3.376538	0.475563
max	57.751384	4.887847	4.458785



(a) Respect to race (b) Respect to sex

Figure A.7: Density and scatter plots for pairwise feature relationships in the Law school admission dataset

Table A.8: Feature statistics for the COMPAS recidivism risk dataset

	JuvFel	JuvMisd	Priors	Crime	Risk	Age
mean	0.055422	0.092290	3.237621	0.354985	4.505135	34.437173
std	0.411727	0.515299	4.920197	0.116691	2.796403	11.753856
min	-1.689539	-2.109104	-16.039693	0.056744	-7.993525	18.000000
25%	-0.225370	-0.255570	-0.089115	0.268702	2.637608	25.000000
50%	0.054229	0.093318	3.241141	0.345316	4.512699	31.000000
75%	0.331007	0.440908	6.562290	0.432243	6.399304	42.000000
max	1.795535	2.194800	22.396505	0.832206	18.477035	80.000000

Figure A.8: Density and scatter plots of pairwise feature relations for the semi-synthetic COMPAS recidivism risk dataset