Personalized Public Policy Analysis in Social Sciences Using Causal-Graphical Normalizing Flows

Downloaded from: https://research.chalmers.se, 2026-09-12 23:25 UTC

Citation for the original published paper (version of record):

Balgi, S., Peña, J., Daoud, A. (2022). Personalized Public Policy Analysis in Social Sciences Using Causal-Graphical Normalizing Flows. Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022, 36: 11810-11818. http://dx.doi.org/10.1609/aaai.v36i11.21437

N.B. When citing this work, cite the original published paper.

research.chalmers.se offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all kind of research output: articles, dissertations, conference papers, reports etc. since 2004. research.chalmers.se is administrated and maintained by Chalmers Library

(article starts on next page)

Personalized Public Policy Analysis in Social Sciences Using Causal-Graphical Normalizing Flows

Sourabh Balgi[1], **Jose M. Peña**[1], **Adel Daoud**[2,3]

[1] The Department of Computer and Information Science (IDA), Linköping University, Sweden
[2] The Institute for Analytical Sociology (IAS), Linköping University, Sweden
[3] The Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden
sourabh.balgi@liu.se, jose.m.pena@liu.se, adel.daoud@liu.se

Abstract

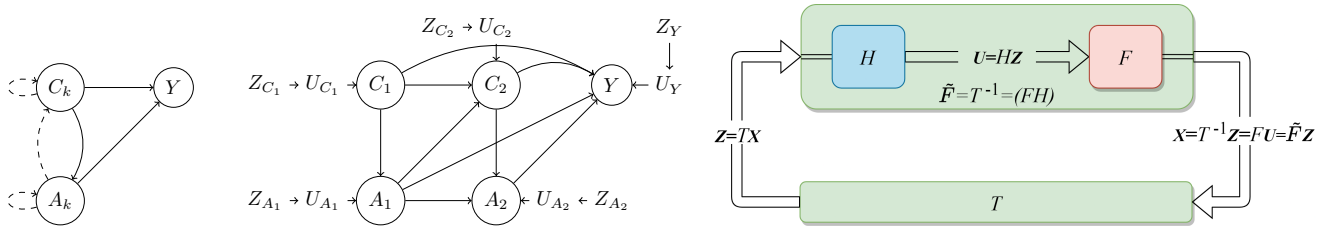
Structural Equation/Causal Models (SEMs/SCMs) are widely used in epidemiology and social sciences to identify and analyze the average causal effect (ACE) and conditional ACE (CACE). Traditional causal effect estimation methods such as Inverse Probability Weighting (IPW) and more recently Regression-With-Residuals (RWR) are widely used - as they avoid the challenging task of identifying the SCM parameters - to estimate ACE and CACE. However, much work remains before traditional estimation methods can be used for counterfactual inference, and for the benefit of Personalized Public Policy Analysis (P3A) in the social sciences. While doctors rely on personalized medicine to tailor treatments to patients in laboratory settings (relatively closed systems), P3A draws inspiration from such tailoring but adapts it for open social systems. In this article, we develop a method for counterfactual inference that we name causal-Graphical Normalizing Flow (c-GNF), facilitating P3A. A major advantage of c-GNF is that it suits the open system in which P3A is conducted. First, we show how c-GNF captures the underlying SCM without making any assumption about functional forms. This capturing capability is enabled by the deep neural networks that model the underlying SCM via observational data likelihood maximization using gradient descent. Second, we propose a novel dequantization trick to deal with discrete variables, which is a limitation of normalizing flows in general. Third, we demonstrate in experiments that c-GNF performs on-par with IPW and RWR in terms of bias and variance for estimating the ACE, when the true functional forms are known, and better when they are unknown. Fourth and most importantly, we conduct counterfactual inference with c-GNFs, demonstrating promising empirical performance. Because IPW and RWR, like other traditional methods, lack the capability of counterfactual inference, c-GNFs will likely play a major role in tailoring personalized treatment, facilitating P3A, optimizing social interventions - in contrast to the current 'one-size-fits-all' approach of existing methods.

Introduction

Since the realization that correlation does not imply causation (Wright 1921; Fisher 1936), statisticians, computer scientists, and social scientists have been developing ways to identify and estimate causal effects from observational data. By the 1970s, Structural Equation Models (SEMs), developed from Wright's path analysis (Wright 1921) in genetics, have been widely used in economics (Haavelmo 1943; Goldberger 1972) and other social sciences (King 1974; Ploch, Goldberger, and Duncan 1975; Fienberg and Duncan 1975) to analyze cause and effect. Pearl's Structural Causal Model (SCM) (Pearl 2009b) provides an explicit causal definition to disambiguate the causes and effects by replacing the symmetric '=' operator with asymmetric ':=' assignment operator, thereby complementing the original SEM definition. Energized by this and similar developments in computer science, a causal revolution has occurred with large ramifications not only in academia but also for how governments, non-governmental organizations, and international organizations (such as the World Bank, United Nations Children's Fund (UNICEF), or International Monetary Fund) articulate public policies to combat social ills. For example, to combat poverty, governments use mainly an individual's income and similar characteristics to identify vulnerable population and then assess if they are eligible for social welfare eligibility. Then a simple rule is often applied: if they are eligible, they tend to receive a fixed one-size-fits-all public policy; if they are not eligible, no policy is ascribed. Although such public-policy making is highly transparent as it applies what is best on average for a population – that is the average causal effect (ACE) – it lacks an adaptability necessary to efficiently combat poverty, ill-health, and other social ills. For effective combating, government officials and others require methods that are able to personalize these policies (Kino et al. 2021). That is, personalized public policy analysis (P^3A) requires methods that can move beyond ACE estimation, and into counterfactual inference.

However, traditional statistical causal effect estimation methods, such as Inverse Probability Weighting (IPW) (Rosenbaum and Rubin 1983; Hernán and Robins 2009) or recent ones such as Regression-With-Residuals (RWR) (Wodtke 2020), focus on ACE estimation only. An advantage of IPW (stabilized) is that it provides a simple and effective way to estimate ACE without the need to model the entire causal system, under the assumption that the functional form of the propensity score is correctly specified. In contrast to IPW, RWR, which comes under the class of outcome regression models, estimate ACEs more effectively (less variance), under the assumption that the func-

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(a) Temporal mechanism. (b) 2-wave model. (c) c-GNF architecture for causal and counterfactual inference.

Figure 1: Fig. 1a shows temporal interactions between the treatments A_k and the observed covariates/confounders C_k at time-wave $k{=}1, \ldots, K$ and the outcome of interest Y. The solid edges indicate the cause and effect at the same time-wave and the dashed edges indicate the one time-wave delayed effect with respect to the respective causes. Fig. 1b represents the 2-wave model of Fig. 1a. For any given observed endogenous variable X in Fig. 1b, U_X and Z_X respectively denote the unobserved exogenous noises of the true SCM $F{:}\mathbf{U}{\to}\mathbf{X}$ and the *encapsulated-SCM* $\tilde{F}{=}(FH){=}T^{-1}{:}\mathbf{Z}{\to}\mathbf{X}$ in Fig. 1c, where $H{:}\mathbf{Z}{\to}\mathbf{U}$ denotes an auxiliary transformation. Since $\tilde{F}$ (green) encapsulates F (red) and H (blue) SCMs, i.e., $\tilde{F}{=}(FH){=}T^{-1}$, we refer to $\tilde{F}$ as the *encapsulated-SCM*. Our c-GNF models $T{:}\mathbf{X}{\to}\mathbf{Z}$ and readily provides the *encapsulated-SCM* $\tilde{F}{:}\mathbf{Z}{\to}\mathbf{X}$, as T is invertible by construction, facilitating counterfactual inference using *'The First Law of Causal Inference'* (Pearl 2009a,b).

tional form of the outcome models are correctly specified. However due to the use of a single outcome model similar to S(Single)-Learner, RWR provides biased estimates when there is effect modification in data (also know as effect heterogeneity) (VanderWeele 2009b).

To directly address effect modification, computer scientists have developed a class of outcome regression based machine-learning models called meta-learners (Künzel et al. 2019) (e.g., T(Two)-Learner, X-Learner) where each treatment group is modeled using a separate outcome regression models conditioned on proper adjustment sets to account for any confounding. Since meta-learners model separate conditional outcome regression models for each treatment groups, we alternatively refer them as Grouped Conditional Outcome Model (GCOM). The conditioning set or adjustment set or the predictors of outcome regression model are identified from the non-parametric expression of the interventional distribution obtained from *do*-calculus (Pearl 2009b, 2012) or G-computation formula (Robins 1986; Hernán and Robins 2009) to accurately model and estimate the causal effects of interest by adjusting for the confounding. While IPW requires the right propensity score model, RWR and GCOM methods require the correct outcome model. In contrast, Doubly Robust (DR) (Bang and Robins 2005) methods are a class of causal inference methods that incorporate the best of IPW and GCOM by jointly modeling the propensity score and the outcome. The key advantage of DR is that it requires only either one of the propensity or the outcome model to be correctly specified to produce unbiased ACE estimates. Notwithstanding the advantages of IPW, RWR, GCOM, and DR, none of them successfully supply a method to conduct counterfactual inference for P^{3}A.

In this work, we aim to further P^{3}A and causal inference in the social science by demonstrating a method for counterfactual inference that does not require a priori knowledge of the correct functional form in neither the propensity score or the outcome model. Our causal method consists of a flow-based deep learning approach known as Graphical Normalizing Flows (GNFs). As we re-purpose GNFs for causal inference by using causal-Directed Acyclic Graphs (causal-DAGs), we refer them as causal-GNFs (c-GNFs). Even though we focus on the social effect of neighborhood on children's education outcomes (Wodtke 2020) as an example of P^{3}A in this article, our contributions are applicable to other settings in epidemiology, economics, sociology, and political science, where counterfactual policy-making is critical to promote social, material, and health outcomes. Our work supplies at least the following five contributions.

1. We show that c-GNF models the underlying SCM of interest in the form of an *encapsulated-SCM* using only the observational data and without making assumptions on the unobserved exogenous noises or the functional mechanisms of the true SCM.

2. Unlike IPW and RWR that require correct functional forms to provide unbiased ACEs, we show that c-GNF needs no functional form assumptions to achieve unbiasedness because the encapsulated-SCM, parameterized by a deep neural networks, has the ability to model any functional form from the observational data via likelihood maximization by gradient descent.

3. We demonstrate experimentally that c-GNF is on-par with IPW and RWR in terms of the bias and variance of the ACE estimates when the right functional forms are assumed and above-par when they are unknown under both small and large sample sizes.

4. Normalizing flows (NFs) are generally built for continuous variables. As this challenges the use of discrete variables, we propose a novel yet simple *Gaussian dequantization* trick to adapt discrete variables for NFs.

5. The most important, unique and key contribution of our c-GNF is that it has the capacity to perform counterfactual inference since c-GNF models the encapsulation of the true SCM. This enables scholars and policymakers to move beyond merely identifying the optimal treatment (public policy) at the population level and start tailoring individual-level optimal treatment (public policy). This capacity will

advance P^{3}A in social science and other domains (such as personalized medicine) (Kino et al. 2021).

In the next section, we introduce the notations, define a typical social problem, and discuss the assumptions required to solve it. Then, we introduce c-GNFs, how they can be used to model SCMs, and how they estimate ACEs and counterfactual inference. Next, we present the simulated experimental setup, results and analysis. Lastly, we conclude our work with implications in social science and beyond.

Notation, Problem Definition and Assumptions

In this section, we discuss the formulation of the causal problem particularly in social science, notations, and critical assumptions for causal inference. Fig. 1a denotes the temporal model involving time-varying confounders used by Wodtke, Harding, and Elwert (2011) in their real-world study involving 16 waves or time-points, i.e., k=1, . . . , 16, where A_k and C_k denote the neighborhood context treatment and the observed neighborhood covariate at time k, and Y denotes the high-school graduation as outcome of causal interest on the Panel Study of Income Dynamics (PSID) sensitive dataset (GeoLytics 2003). However, in this work, we consider the 2-wave model in Fig. 1b used by Wodtke (2020) in his simulated experiments (non-sensitive data) such that the true causal effects are known for benchmarking of different causal effect estimation methods.

Specifically, our aim is same as that of Wodtke (2020), i.e., to estimate the ACEs of the binary treatment variables A_1 and A_2 from the observational data. These ACEs (λ_{a_1,a_2}) are formally defined as

$$\lambda_{1,0} = \mathbf{E}[Y_{1,0}] - \mathbf{E}[Y_{0,0}] \tag{1a}$$

$$\lambda_{0,1} = \mathbf{E}[Y_{0,1}] - \mathbf{E}[Y_{0,0}] \tag{1b}$$

$$\lambda_{1,1} = \mathbf{E}[Y_{1,1}] - \mathbf{E}[Y_{1,0}] \tag{1c}$$

where Y_{a_1,a_2} denotes the potential outcome of Y under the interventions $A_1{:=}a_1$ and $A_2{:=}a_2$. Note that the previous expectations are with respect to the interventional distribution of Y_{a_1,a_2} under the interventions $A_1{:=}a_1$ and $A_2{:=}a_2$

$$P(Y_{\{a_k\}_{k=1}^2}) \ . \tag{2a}$$

Under the assumptions we discuss next, the interventional distribution in Eq. (2a) can be expressed in terms of the components from the observational joint distribution via *do*-calculus (Pearl 2012) or G-computation formula (Robins 1986; Hernán and Robins 2009; Pearl 2009b) as

$$\begin{aligned} P(Y_{\{a_k\}_{k=1}^2}) = \sum_{\{C_k\}_{k=1}^2} & P(Y|\{C_k, a_k\}_{k=1}^2) \\ & P(C_2|C_1, a_1) \\ & \left(\Pi_{j=1}^2 1_{(A_j=a_j)}\right) P(C_1) \ , \end{aligned} \tag{2b}$$

where $1_{(A_j=a_j)}$ is the indicator function denoting the intervention $A_j{:=}a_j$. For the curious readers, the interventional distribution for the general K-wave model of Wodtke, Harding, and Elwert (2011) can be expressed analogously as

$$\begin{aligned} P(Y_{\{a_k\}_{k=1}^K}) = \sum_{\{C_k\}_{k=1}^K} & P(Y|\{C_k, a_k\}_{k=1}^K) \\ & \left(\Pi_{j=2}^K P(C_j|C_{j-1}, a_{j-1})\right) \\ & \left(\Pi_{j=1}^K 1_{(A_j=a_j)}\right) P(C_1) \ . \end{aligned} \tag{2c}$$

Critical Assumptions in Causal Inference

Although generally 'correlation does not imply causation', causal inference is about defining under what assumptions and conditions correlation coincides with causation (Pearl 2009b; Neal 2020). For our counterfactual-inference argument, we require six assumptions and conditions that are briefly stated below. The technical appendix shows detailing of the mathematical definitions and how causal quantities can be estimated from purely statistical quantities.

1. **Unconfoundedness or no unobserved confounders**: As shown in Fig. 1a, we assume that there is no unobserved confounding between A_k and Y. Otherwise, it is impossible to know if the observed correlation is due to causality or confounding. This assumption is strong and untestable (Rubin 1990).

2. **Positivity or overlap or common support or extrapolation**: We assume that every individual has non-zero probability of receiving any of the treatments, i.e., $P(A_k|A_{k-1}, C_k)$ is always positive. Otherwise, it is impossible to accurately estimate the causal effect as the model is inaccurate in the non-overlap region due to no data.

3. **Conditional ignorability or exchangability**: In economics and epidemiology, the assumptions of unconfoundedness and positivity are referred together as conditional ignorability or exchangability (Rosenbaum and Rubin 1983). Unconditional ignorability or exchangability can only be achieved in Randomized Controlled Trials (RCTs).

4. **Consistency**: This assumption states that the observed outcome is same as the potential outcome under the observed treatment (Robins 1986; Cole and Frangakis 2009; VanderWeele 2009a).

5. **No interference or Stable Unit-Treatment Value Assumption (SUTVA)**: No interference means that a particular individual's outcome is only a function of that particular individual's treatment and is not affected by the treatment of any other individual (Cox 1958).

6. **Modularity or independent mechanisms or autonomy or invariance**: Any joint distribution can be factorized as the product of conditional distributions corresponding to independent causal mechanisms using the Markov assumption. Under the interventions, the modularity assumption states that all the mechanisms of the non-intervened variables remains the same while the mechanism of the intervened variables is set to the intervened value. The modularity assumption can be seen in Eqs. (2b) and (2c), where the probability terms corresponding to the mechanism of the intervened variable $A_j{:=}a_j$, i.e., $P(A_j|C_j, A_{j-1})$ term in the observational joint distribution is replaced by the indicator function $1_{(A_j=a_j)}$ denoting the intervention $A_j{:=}a_j$. This modularity assumption is crucial for causal and counterfactual inference (Pearl 2009a,b; Pearl and Mackenzie 2018).

SCM, Encapsulated-SCM and c-GNF

In this section, we present our c-GNFs and describe their connection to SCMs. An SCM consists of a set of assignment equations describing the causal relations between the random variables of a causal system, such as

$$X_i := f_i(PA_i, U_i) \text{ with } i{=}1, \ldots, d \tag{3}$$

where $\{X_i\}_{i=1}^d$ represents the set of observed endogenous random variables, PA_i represents the set of variables that are connoting parents of X_i, $\{U_i\}_{i=1}^d$ denotes the set of exogenous noise random variables, and $\{f_i\}_{i=1}^d$ denotes the set of functions (*independent mechanisms*) that generate the endogenous variable X_i from its observed causes PA_i and noise U_i. SCM is also referred as Functional Causal Model (FCM) due to the functional mechanisms $\{f_i\}_{i=1}^d$.

Let $\mathbf{X}{=}[X_1, \ldots, X_d]^\mathrm{T}{\in}\mathbb{R}^d$ denote the d-dimensional vector of SCM endogenous variables. Similarly, let $\mathbf{U}{=}[U_1, \ldots, U_d]^\mathrm{T}{\in}\mathbb{R}^d$ denote the d-dimensional vector of SCM exogenous variables and $F{:}\mathbf{U}{\to}\mathbf{X}$ denote the transformation representing the true SCM such that $\mathbf{X}{=}F\mathbf{U}$. Let $\mathcal{G}$ represent a causal-DAG with $\{X_1, \ldots, X_d\}$ as the set of nodes and adjacency matrix $\mathcal{A}_\mathcal{G}{\in}\{0,1\}^{d\times d}$. Let $P_\mathbf{X}(\mathbf{X})$ represent the joint distribution over the endogenous variables $\mathbf{X}$, which factorizes according to the causal-DAG $\mathcal{G}$ as

$$P_\mathbf{X}(\mathbf{X}) = \Pi_{i=1}^d P(X_i|PA_i) \ , \tag{4a}$$

where $PA_i{=}\{X_j{:}\mathcal{A}_{\mathcal{G}i,j}{=}1\}$ denotes the set of parents of the vertex/node X_i in the causal-DAG $\mathcal{G}$. As we show in the next section, the factorized joint distribution in Eq. (4a) can be modeled using an autoregressive model parameterized by a deep-neural-network θ, which we denote by $P_\mathbf{X}(\mathbf{X};\theta)$. Such a model is what we call c-GNF.

Causal-Graphical Normalizing Flows (c-GNFs)

Normalizing Flows (NFs) (Tabak and Vanden-Eijnden 2010; Tabak and Turner 2013; Rezende and Mohamed 2015; Kobyzev, Prince, and Brubaker 2020; Papamakarios et al. 2021) are a family of generative models with tractable distributions where both sampling and density evaluation can be efficient and exact. A NF is a flow-based model with transformation $T{:}\mathbf{X}{\to}\mathbf{Z}$, such that $\mathbf{Z}{=}T\mathbf{X}$, where $\mathbf{Z}{=}[Z_1, \ldots, Z_d]^\mathrm{T}{\in}\mathbb{R}^d$ represents the *base random variable* of the flow-model with the *base distribution* $P_\mathbf{Z}(\mathbf{Z})$ that is usually a d-dimensional standard normal distribution for computational convenience and ease of density estimation. Hence the name *normalizing flow*. The defining properties of T are, (i) T must be invertible with T^{-1} as the inverse generative flow such that $T^{-1}{:}\mathbf{Z}{\to}\mathbf{X}$, and (ii) T and T^{-1} must be differentiable, i.e., T must be a d-dimensional diffeomorphism (Milnor and Weaver 1997). Under these properties, from the change of variables formula, we can express the endogenous joint distribution $P_\mathbf{X}(\mathbf{X})$ in terms of the assumed base distribution $P_\mathbf{Z}(\mathbf{Z})$ as

$$P_\mathbf{X}(\mathbf{X}){=}P_\mathbf{Z}(T\mathbf{X})|\mathrm{det}J_T(\mathbf{X})| \ . \tag{4b}$$

Since calculating the joint density of $P_\mathbf{X}(\mathbf{X})$ requires the calculation of the determinant of the Jacobian of T with respect to $\mathbf{X}$, i.e., $\mathrm{det}J_T(\mathbf{X})$, it is advantageous for computational reasons to choose T to have an autoregressive structure such that $J_T(\mathbf{X})$ is a lower-triangular matrix and $\mathrm{det}J_T(\mathbf{X})$ is just the product of the diagonal elements. Autoregressive Flows (AFs) (Kingma et al. 2016; Papamakarios, Murray, and Pavlakou 2017; Huang et al. 2018) are NFs that model autoregressive structure. AFs are composed of two components, the transformer and the conditioner (Papamakarios et al. 2021). Under the assumption of a strictly monotonic transformer, AFs are *universal density estimators* (Huang et al. 2018). Of all the current AFs, Graphical Normalizing Flows (GNFs) (Wehenkel and Louppe 2021) facilitate the use a desired DAG representation as opposed to an arbitrary autoregressive structure. For causal inference, it is crucial that the DAG has a causal interpretation. In this work, we assume the true causal-DAG $\mathcal{G}$ for the GNFs, hence the term causal-GNFs (c-GNFs). In c-GNFs, we use Unconstrained Monotonic Neural Network (UMNN) (Wehenkel and Louppe 2019), a strictly monotonic integration based transformer with $\mathcal{G}$ for the graphical conditioner.

Since Markov equivalent DAGs induce equivalent factorizations of the observational joint distributions, GNFs that use Markov equivalent DAGs represent the same observational joint distribution. This is problematic for causal and counterfactual inference, as different GNFs may represent different interventional joint distributions. In other words, to perform causal inference as we do in this work, the GNF needs to use causal-DAG $\mathcal{G}$. Hence the name causal-GNF (c-GNF). To appreciate this, consider the Causal Autoregressive Flow (CAREFL) by Khemakhem et al. (2021). While CAREFL uses AF, it is flawed as it considers all the predecessors in the topological ordering, i.e., a Markov equivalent DAG, when it should strictly be just the connoting parents/-causes of the given node, i.e., the causal-DAG $\mathcal{G}$. We present an example of CAREFL's flaw in our technical appendix, which shows the conditions under which CAREFL fails in counterfactual inference. Note that only a GNF with the true causal-DAG $\mathcal{G}$ encapsulates the true SCM, thus satisfying the modularity assumption, which is necessary for correct causal and counterfactual inference. This is formalized as

$$\mathbf{X}{=}F\mathbf{U}{=}F(H(\mathbf{Z})){=}(FH)\mathbf{Z}{=}\tilde{F}\mathbf{Z}{=}T^{-1}\mathbf{Z} \ , \tag{5}$$

where $H{:}\mathbf{Z}{\to}\mathbf{U}$ is an auxiliary transformation such that $U_i{=}h_i(Z_i)$ for $i{=}1, \ldots, d$ and $H{=}[h_1, \ldots, h_d]^\mathrm{T}$. It follows from Eq. (5) that our c-GNF $T^{-1}{=}\tilde{F}{=}(FH)$ encapsulates the true SCM F as T and T^{-1} both encode causal-DAG $\mathcal{G}$ in the graphical conditioner, thus providing a way to indirectly model F without making assumptions on $\mathbf{U}$ or the functional causal mechanisms F or the auxiliary transformation H. This sets apart our c-GNFs from most other models for causal inference. Fig. 1c shows the c-GNF architecture for causal and counterfactual inference using T and T^{-1}.

Training a c-GNF amounts to training the deep neural networks that parameterize the transformers and conditioners. This is typically done by maximizing the log-likelihood of the training dataset $\{\mathbf{X}^\ell\}_{\ell=1}^{N_s}$, which is expressed as shown below, by using Eq. (4b) for the summation term $P_\mathbf{X}(\mathbf{X}^\ell;\theta)$

$$\mathcal{L}(\theta){=}\sum_{\ell=1}^{N_s} \log P_\mathbf{X}(\mathbf{X}^\ell;\theta) \tag{6}$$

where θ denotes the parameters of the deep neural networks of the UMNN transformer and graphical conditioner, optimized using stochastic gradient descent.

Gaussian Dequantization Trick NFs naturally model continuous variables, yet in practice, social scientists and others have to model both continuous and discrete variables, as treatments may be discrete categorical variables.

Algorithm 1: Gaussian Dequantization and Quantization

procedure GAUSSIAN DEQUANTIZATION($\{D^\ell\}_{\ell=1}^b$)
  Generate $\tilde{D}^\ell \sim \mathcal{N}(\mu = D^\ell, \sigma^2 = 1/36)$
  return Dequantized / continuous variables $\{\tilde{D}^\ell\}_{\ell=1}^b$
end procedure
procedure QUANTIZATION($\{\tilde{D}^\ell\}_{\ell=1}^b$)
  $D^\ell = \text{clamp}(\text{round}(\tilde{D}^\ell), \text{min}{=}0, \text{max}{=}N{-}1)$
  return Quantized / discrete variables $\{D^\ell\}_{\ell=1}^b$
end procedure

Recently, discrete NFs have been an active field of research including techniques ranging from simple uniform dequantization, Gumbel-max dequantization to complex variational bound dequantization (Uria, Murray, and Larochelle 2013; Hoogeboom et al. 2019; Tran et al. 2019; Ho et al. 2019; Ziegler and Rush 2019; Ma et al. 2019; Nielsen and Winther 2020; Pawlowski, de Castro, and Glocker 2020). Motivated by the Dirac-delta function from control theory, we propose and validate our novel Gaussian dequantization trick (see Algorithm 1 for a discrete variable D with N classes/-categories) to model discrete variables into NFs using the fact that NFs are strongest in modeling Gaussian distributions seamlessly. We provide the exact motivation for the Gaussian dequantization and other details in the technical appendix.

Experiments, Results and Discussion

For our experiments and benchmarking, we simulate a virtual social system using the 2-wave model shown in Fig. 1b. We generated continuous covariates (C_1, C_2), binary treatments (A_1, A_2) and a continuous outcome Y. To provide a realistic sense to the virtual social system under our study, (C_1, C_2) can denote the neighborhood contexts, e.g., the parental income of the individual in the given neighborhood, and (A_1, A_2) can represents neighborhood exposures, e.g., rich/poor neighborhood, and Y can represent the total score obtained in high-school graduation examination. Our aim is to analyse the neighborhood effects on the high-school graduation. We use the following governing equations from Wodtke (2020) to simulate the dataset.

$$C_1 \sim \mathcal{N}(\mu{=}0, \sigma^2{=}1) \ , \tag{7a}$$
$$A_1 \sim Bern(p{=}\Phi(0.4C_1 + 2\gamma_{12}C_1^2)) \ , \tag{7b}$$
$$C_2 \sim \mathcal{N}(\mu{=}0.4C_1 + 0.2A_1, \sigma^2{=}1) \ , \tag{7c}$$
$$A_2 \sim Bern(p{=}\Phi(0.2A_1 + 0.4C_2 + \gamma_{12}A_1^2 + 2\gamma_{12}C_2 + \gamma_{12}C_1A_1/2)) \ , \tag{7d}$$
$$Y \sim \mathcal{N}(\mu{=}0.4(C_1{-}\mu_{C_1}) + A_1(0.2 + \theta_{11}C_1) + (C_2{-}\mu_{C_2})(0.4 + \gamma_{21}C_1) + A_2(0.2 + \theta_{21}C_1 + 0.1A_1), \sigma^2{=}1) \ , \tag{7e}$$

where Φ is the standard normal cumulative distribution function, $\{\theta_{11}, \theta_{21}\}$ are the parameters used to modify the magnitude of the causal effect modification/heterogeneity, and $\{\gamma_{12}, \gamma_{21}\}$ are the parameter used to control the misspecification of the treatment and outcome models.

In the work by Wodtke (2020), every experimental setting is run for 10000 randomly seeded simulations with [500, 1000, 2000] training samples (typical for social science experiments) in each simulation to report the mean ($\mu_{\lambda_{a_1,a_2}}$) and standard-deviation ($\sigma_{\lambda_{a_1,a_2}}$). However, in most practical cases, it is atypical to run deep neural networks for 10000 simulations and is limited between one to 10 in most practical deep learning applications. Hence, we run our experiments with different samples sizes for five randomly seeded simulations, using all the methods (IPW, RWR, GCOM and c-GNF) and report the results pictorially in Fig. 2. Since we run multiple simulations with multiple setting with c-GNFs, we do not concentrate on hyperparameter selection that is best for each simulation and simply use three fully-connected layers with [20, 15, 10] hidden units for the graphical conditioner and three fully-connected layers with [15, 10, 5] hidden units for the monotonic UMNN transformer. We implement c-GNFs in Pytorch (Paszke et al. 2017) using GNF baseline code[1] and AdamW (Loshchilov and Hutter 2019) optimizer with learning-rate=$1e{-}3$ and a batch-size of 128 (2GB of GPU memory) for all our experiments. It is also our aim to evaluate the robustness of c-GNFs in estimating the ACEs without any fine hyperparameters selection using only the most basic/default settings. Since, deep neural networks are prone to overfit, we split our data into three train, validation and test sets in ratio 8:1:1. We strictly use only the training set with [500, 2000, 4000000] samples for training and use the held-out validation set for early stopping to get the model with best validation loss. We further validate the generalization of the best validation loss model on the held-out test set. We simulate 2000 samples from every interventional distributions using c-GNFs to evaluate $\mathbf{E}[Y_{a_1,a_2}]$ to obtain Monte-Carlo estimates of ACEs using Eqs. (1a)-(1c). Theoretically, a higher number of samples from the interventional distributions would provide better Monte-Carlo estimates at an added computational cost.

Fig. 2 indicates the results for the following three experimental settings: (i) Fig. 2a shows the results with both the treatment and outcome models correctly specified without heterogeneity in data. (ii) Fig. 2b shows the results with only the treatment model correctly specified with heterogeneity in data. (iii) Fig. 2c shows the results with both treatment and outcome models misspecified with heterogeneity in data. Under all three experimental settings, the true ACEs of interest for the above 2-wave time-varying model λ_{a_1,a_2} for the binary treatments $\{A_1{:=}a_1, A_2{:=}a_2\}$ are calculated from Eqs. (1b)-(1c) as $\lambda_{1,0}{=}0.2{-}0{=}0.2$, $\lambda_{0,1}{=}0.2{-}0{=}0.2$, and $\lambda_{1,1}{=}0.5{-}0.2{=}0.3$ and denoted by green horizontal dotted lines in Fig. 2. The red horizontal dotted line in Fig. 2 represents *zero*-ACE which can be considered as the critical point due to the fact that *zero*-ACE indicates both the treatments to have the same causal effect/outcome. Sometimes, it is more important to accurately identify the sign of the ACE with some acceptable tolerance in the magnitude of the ACE estimate as the sign is what determines the effectiveness of the treatment, e.g., $\lambda_{1,0}{>}0$ in Eq. (1a) indicates treatment $\{A_1{:=}1, A_2{:=}0\}$ is better than treatment $\{A_1{:=}0, A_2{:=}0\}$

[1] https://github.com/AWehenkel/Graphical-Normalizing-Flows

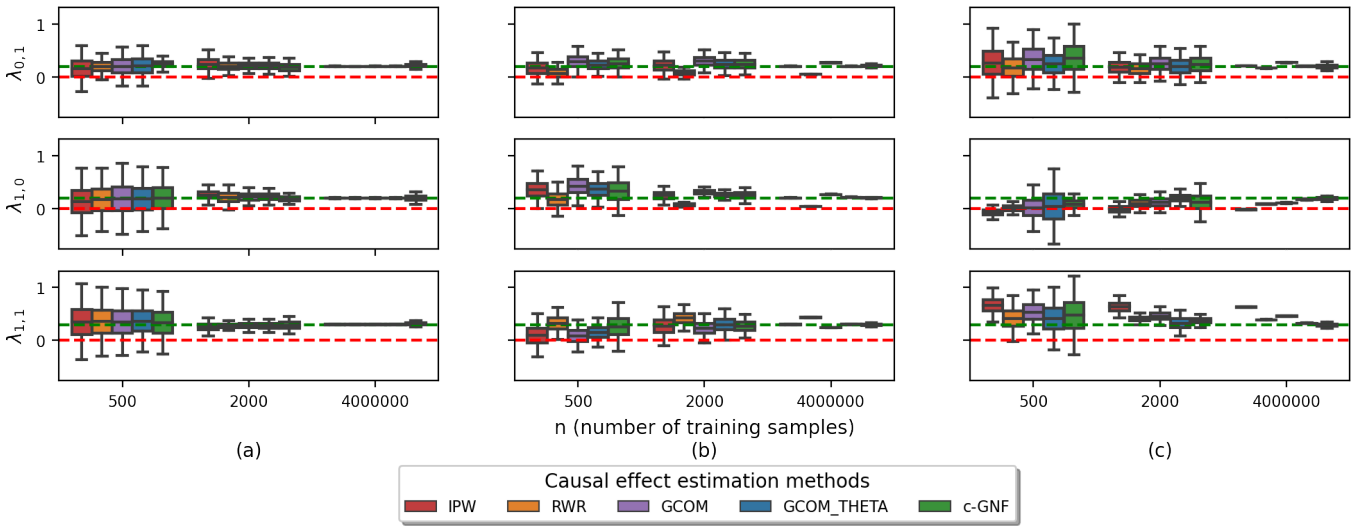


Figure 2: (a) $\{\theta_{11}{=}\theta_{21}{=}0, \gamma_{21}{=}\gamma_{12}{=}0\}$. (b) $\{\theta_{11}{=}\theta_{21}{=}0.2, \gamma_{21}{=}0.4, \gamma_{12}{=}0\}$. (c) $\{\theta_{11}{=}\theta_{21}{=}0.2, \gamma_{21}{=}\gamma_{12}{=}0.4\}$.

in expectation, and vice-versa if $\lambda_{1,0}{<}0$. It is of primary importance to identify which treatment is better and of secondary importance to identify by what amount the treatment is better, in expectation. In Fig. 2, the box-plots indicate $\mu, \mu{\pm}\sigma, \mu{\pm}3\sigma$ values for all λ_{a_1,a_2} to provide an idea of the variance of the estimates. Observe that in all the plots, c-GNF does not include the *zero*-ACE in $\mu{\pm}\sigma$ range indicating that c-GNF does not introduce ambiguity in the selection of best treatment, in expectation, unlike IPW, RWR and GCOM. This non-ambiguity is a desired quality from any causal effect estimation methods used in decision-making.

In Fig. 2a, we see that all the models perform equally well under no model misspecification and effect modification/heterogeneity in data. However increasing the complexity by adding outcome model misspecification and data heterogeneity in Fig. 2b, we see that the regression based outcome methods not assuming the right functional forms, i.e., GCOM and RWR, result in biased estimates in both small and large samples. GCOM_THETA in Fig. 2 refers correcting GCOM by specifying the true functional form manually or parameterizing the model by a deep neural network as in the case of meta-learners to result in unbiased estimates. Hence, we see that GCOM_THETA performs better than GCOM due to the correction. Similarly, since the treatment model is rightly specified, IPW estimates are unbiased. In contrast to Figs. 2a and 2b, Fig. 2c indicates the treatment model is misspecified as well. With the treatment model misspecified, IPW no longer results in unbiased estimates. From Fig. 2, we show that in small-sample size regime, c-GNF results in unbiased estimates with comparable variance. We can still do better in terms of variance by appropriate selection of the deep neural networks as a part of the hyperparameter tuning, which is not our aim. A real-world application of c-GNF on a large-scale non-randomized observational study to analyse the impact of IMF (International Monetary Fund) program on the child poverty is conducted in Balgi, Peña, and Daoud (2022) that observes an effective reduction of child poverty by $1.2{\pm}0.24$ degrees in the Global-South, thus indicating the beneficial nature of IMF program on child poverty reduction.

Counterfactual Inference

Apart from the benefits of ACE estimates with small bias and variance compared to standard statistical methods, c-GNF provides the unique benefit of counterfactual inference using *'The First Law of Causal Inference'* (Pearl 2009a,b; Pearl and Mackenzie 2018), which is virtually non-existent in IPW/RWR/GCOM. 'The First Law of Causal Inference' essentially provides a framework to identify the unit level potential outcomes using SCM, thereby addressing the fundamental missing value problem of counterfactual inference. This law to identify the missing/unseen potential outcome Y_{a_1,a_2} for a unit $\mathbf{Z}^\ell$ is represented mathematically as

$$Y_{a_1,a_2}(\mathbf{Z}^\ell){=}Y_{\tilde{F}_{a_1,a_2}}(\mathbf{Z}^\ell) \ , \tag{8}$$

where $Y_{a_1,a_2}(\mathbf{Z}^\ell)$ denotes the potential outcome Y under the treatments $\{A_j{:=}a_j\}_{j=1}^2$ for a given unit individual $\mathbf{Z}^\ell$ that is fundamentally missing and is of interest for counterfactual inference, and $Y_{\tilde{F}_{a_1,a_2}}(\mathbf{Z}^\ell)$ represents the actual outcome Y observed from the mutilated-SCM $\tilde{F}_{a_1,a_2}$ that is obtained by the mutilation $\{A_j{:=}a_j\}_{j=1}^2$ on the *encapsulated-SCM* $\tilde{F}$. Essentially, this law involves three steps that help identify the missing potential outcomes, thereby addressing the fundamental missing value problem of causal inference.

1. **Abduction**: For a given observed unit-individual evidence $\mathbf{X}^\ell$, the respective exogenous SCM noise $\mathbf{Z}^\ell$ corresponding to $\mathbf{X}^\ell$ is recovered from the encapsulated-SCM $\tilde{F}$. In case of c-GNF, we have $\mathbf{Z}^\ell{=}T\mathbf{X}^\ell{=}\tilde{F}^{-1}\mathbf{X}^\ell$ where the *encapsulated-SCM* noise $\mathbf{Z}^\ell$ indicates the unique hidden essence/identifier/DNA of the ℓ^{th} unit-individual with observed evidence $\mathbf{X}^\ell$.

2. **Action**: The action or intervention (in our case a public policy) corresponding to the desired treatment is con-

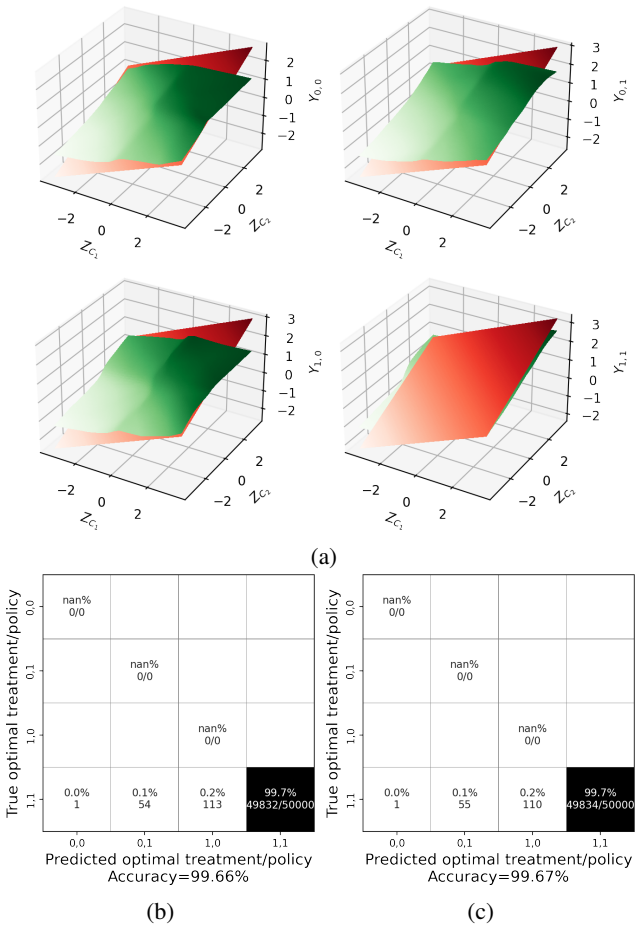


Figure 3: (a) The red and green surface-plots respectively represent the potential outcome maps of true SCM F and our *encapsulated-SCM* $\tilde{F}$ generated from the SCM noise ranges $Z_{C_1} \in [-3, 3], Z_{C_2} \in [-3, 3], Z_Y = 0$. (b, c) Confusion matrices between the true and predicted optimal personalised policy on unseen validation and test sets respectively (with model under no misspecification setting from Fig. 2a).

ducted in the form of mutilating the corresponding structural equation of the treatment $\{A_j := a_j\}_{j=1}^{2}$ in $\tilde{F}$ (modularity assumption) resulting in the mutilated-SCM $\tilde{F}_{a_1,a_2}$.

3. **Prediction**: The recovered noises $\mathbf{Z}^\ell$ from abduction are re-propagated through $\tilde{F}_{a_1,a_2}$ and the outcomes $Y_{\tilde{F}_{a_1,a_2}}(\mathbf{Z}^\ell)$ observed are nothing but the potential outcomes $Y_{a_1,a_2}(\mathbf{Z}^\ell)$ we were interested to begin with.

The potential outcome maps in Fig. 3a are generated using the three steps of counterfactual inference with c-GNF, i.e., *encapsulated-SCM* and the noises $Z_{C_1} \in [-3, 3], Z_{C_2} \in [-3, 3], Z_Y = 0$ to observe the effect of C_1 and C_2 on Y. Due to the linear nature of the structural equation in Eq. (7e) under no model misspecification setting, the true potential outcome maps are represented by the flat hyperplanes in Fig. 3a (red). Observe that around the corners of the green 3D surface-plots which represents the low likelihood region of observing samples, the extrapolation assumption is slightly violated as the observational data corresponding to these regions are likely missing in the training set due to low likelihood. Extrapolation is an inevitable assumption made in almost all statistical models which is almost always violated under practical finite data setting. True extrapolation can be attained only with infinite data, which is impractical and unrealistic. In the technical appendix, we further present the same set of plots as in Fig. 3 under the complex experimental setting when both the models are misspecified and also the data exhibits effect modification/heterogeneity.

The confusion matrices from unseen validation and test sets in Figs. 3b and 3c indicate the true and predicted optimal individual-level treatment/policy $\bar{a}_1^\ell, \bar{a}_2^\ell$ for the ℓ^{th} individual $\mathbf{X}^\ell$, which is predicted as below from Eq. (8)

$$\bar{a}_1^\ell, \bar{a}_2^\ell = \underset{a_1,a_2}{\operatorname{argmax}}\, Y_{a_1,a_2}(T\mathbf{X}^\ell) \ . \tag{9}$$

Using c-GNF, we observe that the optimal treatments are indeed predicted with a very high accuracy of 99.6% on unseen validation and test sets under no-misspecification and no effect modification/heterogeneity in data. Similarly, under both model misspecifications we obtain 91% accuracy even with our sub-optimal hyperparameters. Thus, c-GNF enables precise P^{3}A targeting and articulation of the tailored treatments/policies in the social sciences, instead of relying on 'one-size-fits-all' sub-optimal treatments/policies.

Finally, our c-GNF not only model the encapsulation of the true SCM, but also offer exact density estimation which provides necessary tools for causal inference practitioners to identify other causal quantities such as the probabilities of causation (Pearl 1999, 2009b), i.e., Probability of Necessity (PN), Probability of Sufficiency (PS), Probability of Necessity and Sufficiency (PNS), Probability of Disablement (PD), Probability of Enablement (PE), that are of utmost importance to policy-makers planning various social programs.

Conclusion

This article developed causal-Graphical Normalizing Flow (*c-GNF*) for personalized public policy analysis (P^{3}A). We demonstrated that our c-GNF learnt using only observational data without any assumptions on the exogenous noises or the functional mechanisms of the underlying SCM.We further identified c-GNF is on/above-par well-established standard causal effect estimation methods such as IPW and RWR. Most importantly, in contrast to IPW and RWR, using simulation experiments, we demonstrated the unique benefit of counterfactual inference using *'The First Law of Causal Inference'* with c-GNF. Although our simulation study might be perceived as a limitation, the strong benchmarking of c-GNF on the simulated dataset demonstrated the usability of our framework with real-world data. The counterfactual inference demonstrated with c-GNFs showed potential to identify the optimal treatments at an individual-level, enabling precise P^{3}A. When successful, P^{3}A will likely accelerate the use of individualized policies beyond the closed settings of medical applications and into the open settings of the social science.

Acknowledgments

The authors would like to acknowledge the generous funding from the Swedish Research Council through the Swedish Network for Register-Based Research (SWE-REG) with respect to the grant number 2019-00245 with an aim to advance machine learning to tackle problems in social sciences through register-based-learning using the Swedish population registry.

Ethics Statement

The authors would like to indicate that the current work does not involve any real ethical considerations as assumptions are clearly stated and the experimental results with benchmarking are based only on the non-sensitive simulated dataset.

References

Balgi, S.; Peña, J. M.; and Daoud, A. 2022. Counterfactual Analysis of the Impact of the IMF Program on Child Poverty in the Global-South Region using Causal-Graphical Normalizing Flows. *arXiv preprint arXiv:2202.09391*.

Bang, H.; and Robins, J. M. 2005. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61(4): 962–972.

Cole, S. R.; and Frangakis, C. E. 2009. The Consistency Statement in Causal Inference: A Definition or An Assumption? *Epidemiology*, 20(1): 3–5.

Cox, D. R. 1958. *Planning of Experiments*. New York: Wiley.

Fienberg, S.; and Duncan, O. D. 1975. Introduction to Structural Equation Models. *Journal of the American Statistical Association (JASA)*, 72: 485.

Fisher, R. A. 1936. Design of Experiments. *British Medical Journal (BMJ)*, 1(3923): 554–554.

GeoLytics, I. 2003. CensusCD Neighborhood Change Database, 1970–2000 Tract Data. *New Brunswick, NJ: GeoLytics*.

Goldberger, A. S. 1972. Structural Equation Methods in the Social Sciences. *Econometrica*, 40(6): 979–1001.

Haavelmo, T. 1943. The Statistical Implications of a System of Simultaneous Equations. *Econometrica*, 11: 1–12.

Hernán, M. A.; and Robins, J. M. 2009. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Ho, J.; Chen, X.; Srinivas, A.; Duan, Y.; and Abbeel, P. 2019. Flow++: Improving Flow-Based Generative Models with Variational Dequantization and Architecture Design. In *International Conference on Machine Learning (ICML)*, 2722–2730.

Hoogeboom, E.; Peters, J. W. T.; van den Berg, R.; and Welling, M. 2019. Integer Discrete Flows and Lossless Compression. In *Neural Information Processing Systems (NeurIPS)*, 12134–12144.

Huang, C.; Krueger, D.; Lacoste, A.; and Courville, A. C. 2018. Neural Autoregressive Flows. In *International Conference on Machine Learning (ICML)*, 2083–2092.

Khemakhem, I.; Monti, R.; Leech, R.; and Hyvarinen, A. 2021. Causal Autoregressive Flows. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 3520–3528.

King, M. 1974. A. S. Goldberger and O. D. Duncan. Structural Equation Models in the Social Sciences. *The Economic Journal*, 84(333): 212–214.

Kingma, D. P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; and Welling, M. 2016. Improved Variational Inference with Inverse Autoregressive Flow. In *Neural Information Processing Systems (NeurIPS)*, 4743–4751.

Kino, S.; Hsu, Y.-T.; Shiba, K.; Chien, Y.-S.; Mita, C.; Kawachi, I.; and Daoud, A. 2021. A Scoping Review on the Use of Machine Learning in Research on Social Determinants of Health: Trends and Research Prospects. *Social Science & Medicine - Population Health (SSM-PH)*, 15: 100836–100855.

Kobyzev, I.; Prince, S.; and Brubaker, M. 2020. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1–1.

Künzel, S. R.; Sekhon, J. S.; Bickel, P. J.; and Yu, B. 2019. Metalearners for Estimating Heterogeneous Treatment Effects using Machine Learning. *Proceedings of the National Academy of Sciences (PNAS)*, 116(10): 4156–4165.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*.

Ma, X.; Kong, X.; Zhang, S.; and Hovy, E. H. 2019. MaCow: Masked Convolutional Generative Flow. In *Neural Information Processing Systems (NeurIPS)*, 5891–5900.

Milnor, J.; and Weaver, D. W. 1997. *Topology from the Differentiable Viewpoint*. Princeton university press.

Neal, B. 2020. Introduction to Causal Inference from a Machine Learning Perspective. *Course Lecture Notes (draft)*.

Nielsen, D.; and Winther, O. 2020. Closing the Dequantization Gap: PixelCNN as a Single-Layer Flow. In *Neural Information Processing Systems (NeurIPS)*.

Papamakarios, G.; Murray, I.; and Pavlakou, T. 2017. Masked Autoregressive Flow for Density Estimation. In *Neural Information Processing Systems (NeurIPS)*, 2338–2347.

Papamakarios, G.; Nalisnick, E.; Rezende, D. J.; Mohamed, S.; and Lakshminarayanan, B. 2021. Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research (JMLR)*, 22(57): 1–64.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic Differentiation in PyTorch. *NeurIPS Workshops*.

Pawlowski, N.; de Castro, D. C.; and Glocker, B. 2020. Deep Structural Causal Models for Tractable Counterfactual Inference. In *Neural Information Processing Systems (NeurIPS)*.

Pearl, J. 1999. Probabilities of Causation: Three Counterfactual Interpretations and Their Identification. *Synthese*, 121(1): 93–149.

Pearl, J. 2009a. Causal Inference in Statistics: An Overview. *Statistics Surveys*, 3: 96–146.

Pearl, J. 2009b. *Causality: Models, Reasoning and Inference*. USA: Cambridge University Press, 2nd edition. ISBN 052189560X.

Pearl, J. 2012. The *Do*-Calculus Revisited. In *Uncertainty in Artificial Intelligence (UAI)*, 3–11.

Pearl, J.; and Mackenzie, D. 2018. *The Book of Why: The New Science of Cause and Effect*. USA: Basic Books, Inc., 1st edition. ISBN 046509760X.

Ploch, D. R.; Goldberger, A. S.; and Duncan, O. D. 1975. Structural Equations Models in the Social Sciences. *Social Forces*, 54: 503.

Rezende, D.; and Mohamed, S. 2015. Variational Inference with Normalizing Flows. In *International Conference on Machine Learning (ICML)*, 1530–1538.

Robins, J. M. 1986. A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect. *Mathematical Modelling*, 7(9): 1393–1512.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1): 41–55.

Rubin, D. B. 1990. Formal Mode of Statistical Inference for Causal Effects. *Journal of Statistical Planning and Inference (JSPI)*, 25(3): 279–292.

Tabak, E.; and Turner, C. 2013. A Family of Nonparametric Density Estimation Algorithms. *Communications on Pure and Applied Mathematics*, 66(2): 145–164.

Tabak, E.; and Vanden-Eijnden, E. 2010. Density Estimation by Dual Ascent of the Log-Likelihood. *Communications in Mathematical Sciences*, 8(1): 217–233.

Tran, D.; Vafa, K.; Agrawal, K. K.; Dinh, L.; and Poole, B. 2019. Discrete Flows: Invertible Generative Models of Discrete Data. In *Neural Information Processing Systems (NeurIPS)*, 14692–14701.

Uria, B.; Murray, I.; and Larochelle, H. 2013. RNADE: The Real-Valued Neural Autoregressive Density-Estimator. In *Neural Information Processing Systems (NeurIPS)*, 2175–2183.

VanderWeele, T. J. 2009a. Concerning the Consistency Assumption in Causal Inference. *Epidemiology*, 20(6): 880–883.

VanderWeele, T. J. 2009b. On the Distinction Between Interaction and Effect Modification. *Epidemiology*, 20(6): 863–871.

Wehenkel, A.; and Louppe, G. 2019. Unconstrained Monotonic Neural Networks. In *Neural Information Processing Systems (NeurIPS)*, 1545–1555.

Wehenkel, A.; and Louppe, G. 2021. Graphical Normalizing Flows. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 37–45.

Wodtke, G. T. 2020. Regression-Based Adjustment for Time-Varying Confounders. *Sociological Methods & Research (SMR)*, 49(4): 906–946.

Wodtke, G. T.; Harding, D. J.; and Elwert, F. 2011. Neighborhood Effects in Temporal Perspective: The Impact of Long-Term Exposure to Concentrated Disadvantage on High School Graduation. *American Sociological Review (ASR)*, 76(5): 713–736. PMID: 22879678.

Wright, S. 1921. Correlation and Causation. *Journal of Agricultural Research (JAR)*, 20: 557–585.

Ziegler, Z. M.; and Rush, A. M. 2019. Latent Normalizing Flows for Discrete Sequences. In *International Conference on Machine Learning (ICML)*, 7673–7682.