International Journal of Approximate Reasoning 187 (2025) 109531

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Sensitivity analysis to unobserved confounding with copula-based normalizing flows

Sourabh Balgi [a], Marc Braun [a], Jose M. Peña [a,*], Adel Daoud [b]

[a] IDA, Linköping University, Linköping, 58183, Sweden
[b] IEI, Linköping University, Linköping, 58183, Sweden

ARTICLE INFO

Keywords:
Sensitivity analysis
Unconfoundness
Structural causal model
Normalizing flow
Gaussian copula

ABSTRACT

We propose a novel method for sensitivity analysis to unobserved confounding in causal inference. The method builds on a copula-based causal graphical normalizing flow that we term ρ-GNF, where $\rho \in [-1, +1]$ is the sensitivity parameter. The parameter represents the non-causal association between exposure and outcome due to unobserved confounding, which is modeled as a Gaussian copula. In other words, the ρ-GNF enables scholars to estimate the average causal effect (ACE) as a function of ρ, accounting for various confounding strengths. The output of the ρ-GNF is what we term the ρ_{curve}, which provides the bounds for the ACE given an interval of assumed ρ values. The ρ_{curve} also enables scholars to identify the confounding strength required to nullify the ACE. We also propose a Bayesian version of our sensitivity analysis method. Assuming a prior over the sensitivity parameter ρ enables us to derive the posterior distribution over the ACE, which enables us to derive credible intervals. Finally, leveraging on experiments from simulated and real-world data, we show the benefits of our sensitivity analysis method.

1. Introduction

Epidemiologists, sociologists, economists, and other applied scientists, often leverage randomized controlled trials (RCTs), as these provide the safest methodological route to perform causal inference [1–4]. By randomizing which experimental subjects (e.g., people, villages, schools) should take the treatment and which subjects should abstain, an RCT ensures unconfoundedness (also known as ignorability or exchangeability): There is no unobserved common causes of the treatment and the outcome in the causal system under study. When unconfoundedness is satisfied, scholars can calculate the causal effect of interest from collected data; that means that the causal quantity is identifiable [5,6]. However, despite the importance of RCTs, they remain infeasible for a slew of applied settings. They may be costly to implement (e.g., testing a population-wide medicine); they may be unethical (e.g., testing a new drug); or they may be impractical to implement (e.g., testing a social policy across the world). Therefore, applied researchers often rely on observational data – which are often secondary data sources with no treatment randomization and where the experimenter had no control over the data generating process. Yet when using observational data, scholars make themselves susceptible for failing to satisfy the unconfoundedness assumption, even when some confounders are observed.

Because the unconfoundedness assumption is so critical and at the same time untestable in observational studies [5,6], methodologists (statisticians, computer scientist, and others) have developed various frameworks for stress testing how causal effect estimates change under varying degree of unconfoundedness failure. These sorts of tests are named sensitivity analysis [7–11], also known as

* Corresponding author.
E-mail address: jose.m.pena@liu.se (J.M. Peña).

https://doi.org/10.1016/j.ijar.2025.109531
Received 31 January 2025; Received in revised form 21 June 2025; Accepted 16 July 2025

Available online 30 July 2025

bias analysis in epidemiology [12–15]. Nonetheless, existing sensitivity analysis methods are limited in at least three ways, as we elaborate on below and in the next section. In this paper, we propose a new method to improve on these limitations, thereby moving the state-of-the-art forward.

First, we propose a deep learning method for sensitivity analysis based on causal graphical normalizing flows [16–18], because of their attractive properties of non-linearity and invertibility for counterfactual inference. Specifically, we combine these normalizing flows with the Gaussian copula [19], the most studied and popular elliptical copula, to model unobserved confounding. Hence, we aptly name the new model ρ-GNF, where $\rho \in [-1, +1]$ is the sensitivity parameter of the Gaussian copula that controls the degree of unconfoundedness between treatment and outcome. Unlike most sensitivity analysis methods where the sensitivity parameters are unbounded and thus difficult to specify and interpret, ρ is bounded and has a clear interpretation.

Second, we show that our ρ-GNF enables us to estimate the average causal effect (ACE) as a function of ρ. This function, which we call ρ_{curve}, enables us to identify the ACE bounds given a specific interval of ρ that the domain expert considers appropriate. It also enables us to determine the confounding strength required to explain away the causal effect, which we call ρ_{value}. This is similar to the widely used E-value [20]. Unlike most sensitivity analysis methods, our method can be cast in a Bayesian setting by allowing the user to specify a prior distribution over ρ, to return the posterior distribution over the ACE and corresponding credible intervals.

Third, ρ-GNF accommodates both discrete and continuous outcomes. This is in contrast to most existing sensitivity analysis methods, which apply to only one type of outcome. For instance, the works [21,8,22–24] require the outcome variable being binary, while the works [25,26] require the outcome to be continuous.

The remainder of the paper is structured as follows. After reviewing the related literature in Section 2, we introduce ρ-GNF for sensitivity analysis in Section 3. Moreover, we introduce a Bayesian extension of it in Section 4, and discuss suitable choices for prior distributions of ρ. In Section 5, we present our results with simulated and real-world data under different settings of the outcome variable (i.e., continuous, binary and categorical), and compare them with the popular assumption free bounds [21,8]. Finally, in Section 6, we conclude by discussing the key contributions of our ρ-GNF method.

It is worth mentioning that this work is an extension of the conference contribution [27]. Specifically, the extension consists of the Bayesian framework presented in Section 4, its evaluation on simulated and real-world datasets in Sections 5.1 and 5.3, and the proofs of the formal results included in Appendix A and Appendix B.

2. Related works

The literature on sensitivity analysis can be roughly categorized into two streams: (i) identify the bounds of the causal effect as functions of some sensitivity parameters that encode the strength of the unobserved confounders [21,8,11,28,22–24], and (ii) identify how large the influence of the unobserved confounders must be to explain away the causal effect [9,20,29,30]. For instance, the works [21,8] provide assumption free bounds of the causal effect for binary outcomes, while more recent works extend the bounds to categorical outcomes [31]. Other methods provide bounds as functions of sensitivity parameters to be tuned by a domain expert [22–24]. Others are limited to the linear setting [25,26]. In contrast to the bounds stream, there are methods that fall under the explain-away stream. For example, the works [9,20] reason on the minimum strength of the unmeasured confounder that is needed, conditional on the measured confounders, to explain away the causal effect. Similarly, the recently developed Austen plots identify the influence of the confounding needed to explain a specific amount of bias in the causal effect estimate [29].

As shown above, there is a wide spectrum of sensitivity analysis methods. All certainly have unique advantages, but they also have limitations. For example, the methods in the works [21,8,22–24] require the outcome variable being binary. The methods in the work [25,26] assume a linear model. The method in the work [20] may result in wider bounds than the assumption free bounds [32,22], which are thus logically impossible. Some methods are exclusively suited for specific causal estimands, e.g., ACE or conditional ACE or mediation effects [33,34]. Some methods offer no sensitivity parameters to determine the confounding strength required to explain away the causal effect [21,8]. Others like those in the works [20,29] offer multiple parameters that are unbounded and hard to specify for the domain analyst [32].

To summarize, even though there exist several sensitivity analysis frameworks, there is still a lack of unifying method that is flexible enough to suit many different types of observational data, with easy-to-use sensitivity parameters. Moreover, it is imperative to establish a method enabling researchers to specify distributional assumptions regarding the unobserved causes within the causal system under study. Such assumptions enable tighter ACE bounds, enhancing the certainty of the findings. Our ρ-GNF method targets all these lacks. Our method uses deep neural networks, allowing for maximum flexibility and non-linearity. It provides a single bounded sensitivity parameter $\rho \in [-1, +1]$ that is easily interpreted as the measure of non-causal association between exposure and outcome due to unobserved confounding. Moreover, our method can accommodate Bayesian inference over the ACE bounds. Thus, our method enhances an applied researcher's causal toolbox.

3. Sensitivity analysis with copula-based normalizing flows

3.1. Notation and problem definition

Let us consider the following the structural causal model (SCM) [5,35], where A is the treatment (or exposure or cause) and Y is the outcome (or effect), and ε_A and ε_Y are their respective unobserved causes such that

$$A := t_A(\varepsilon_A) \ , \ Y := t_Y(A, \varepsilon_Y) = t_{Y|A}(\varepsilon_Y) \ , \ (\varepsilon_A, \varepsilon_Y) \sim F_{\varepsilon_A, \varepsilon_Y}(\varepsilon_A, \varepsilon_Y) \tag{1}$$

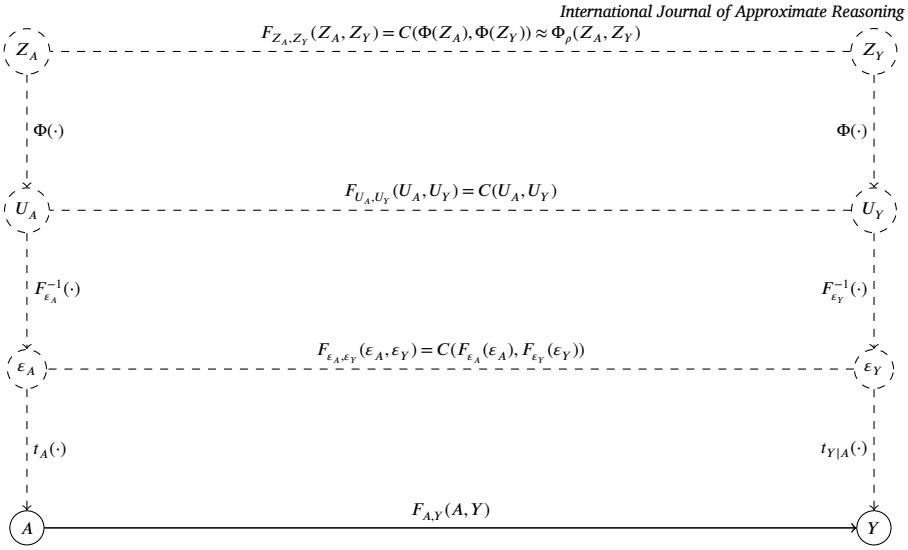


Fig. 1. Graphical representation of Equations (1)-(4).

where t_A and $t_{Y|A}$ are arbitrary transformations, and $(\varepsilon_A, \varepsilon_Y)$ follows an arbitrary joint CDF $F_{\varepsilon_A,\varepsilon_Y}(\varepsilon_A, \varepsilon_Y)$.

Using the universality of the uniform (also known as the probability integral transform) [36], the noise variables ε_A and ε_Y of the SCM in Equation (1) can equivalently be written in terms of uniform variables U_A and U_Y in the interval $[0, 1]$ resulting in the following SCM:

$$A := t_A(F^{-1}_{\varepsilon_A}(U_A)) \ , \ Y := t_{Y|A}(F^{-1}_{\varepsilon_Y}(U_Y)) \ , \ (U_A, U_Y) \sim F_{U_A,U_Y}(U_A, U_Y) \tag{2}$$

where F_{ε_A} and F_{ε_Y} respectively denote the marginal CDFs of ε_A and ε_Y, and (U_A, U_Y) follows the joint CDF $F_{U_A,U_Y}(U_A, U_Y)$ with uniform marginals in [0,1]. From the universality of the uniform, Equation (2) can further be simplified in terms of F_A and $F_{Y|A}$, which respectively denote the marginal CDFs of A and Y conditioned on A:

$$A := F^{-1}_A(U_A) \ , \ Y := F^{-1}_{Y|A}(U_Y) \ , \ (U_A, U_Y) \sim F_{U_A,U_Y}(U_A, U_Y) \ . \tag{3}$$

Furthermore, we can represent U_A and U_Y in Equation (3) as transformations of standard normal variables using the standard normal CDF Φ:

$$A := F^{-1}_A(\Phi(Z_A)) \ , \ Y := F^{-1}_{Y|A}(\Phi(Z_Y)) \ , \ (Z_A, Z_Y) \sim F_{Z_A,Z_Y}(Z_A, Z_Y) \tag{4}$$

where (Z_A, Z_Y) follows the joint CDF $F_{Z_A,Z_Y}(Z_A, Z_Y)$ with standard normal marginals. A graphical representation of Equations (1)-(4) can be found in Fig. 1.

Equations (1)-(4) represent observationally and interventionally equivalent SCMs, but with different unobserved noises and corresponding joint CDFs [5,37]. Observational equivalence means that the SCMs yield the same distribution $F_{A,Y}(A, Y)$. Interventional equivalence means that the SCMs yield the same interventional distributions $F_Y(Y|do(a))$ and $F_A(A|do(y))$. Since these noises are unobserved, the interventional distribution of interest $F_Y(Y|do(a))$ is not identifiable from any of the SCMs, without further assumptions [5]. To overcome this problem, we propose in the subsequent sections to use a bivariate Gaussian copula to model the noise distributions (i.e., confounding strength) so that the interventional distribution of interest can parametrically be estimated using deep-neural-network-inspired normalizing flows trained on observational data.

3.2. Representing confounding with a Gaussian copula

A copula is a multivariate distribution function defined on the unit hypercube with uniform marginals [19]. As the name suggests, a copula ties or links or couples a multidimensional joint distribution to its marginals. Therefore, by Sklar's theorem [19], we have that the bivariate joint CDF $F_{U_A,U_Y}(U_A, U_Y)$ in Equations (2) and (3) with uniform marginals in [0,1] can be represented using some bivariate copula $C(U_A, U_Y)$:

$$F_{U_A,U_Y}(U_A, U_Y) = C(U_A, U_Y) \tag{5}$$

$$F_{\varepsilon_A,\varepsilon_Y}(\varepsilon_A, \varepsilon_Y) = C(F_{\varepsilon_A}(\varepsilon_A), F_{\varepsilon_Y}(\varepsilon_Y)) \tag{6}$$

$$F_{Z_A,Z_Y}(Z_A, Z_Y) = C(\Phi(Z_A), \Phi(Z_Y)) \ . \tag{7}$$

Equations (6) and (7) follow from the scale-invariance property of the copula $C(U_A, U_Y)$ to the strictly increasing transformations F_{ε_A}, F_{ε_Y}, and Φ [19]. The copula essentially models the confounding effect (i.e., non-causal back-door association) between exposure

and outcome in Equations (1)-(4). See also Fig. 1. This may be quantified by measures of association between U_A and U_Y such as Spearman correlation ρ_S [38] or Kendall correlation τ_K [39]. By the scale-invariance property of ρ_S and τ_K to the strictly increasing transformations F_{ε_A}, F_{ε_Y}, and Φ, these measures of association between U_A and U_Y are the same as between ε_A and ε_Y, and between Z_A and Z_Y.

As discussed above, the interventional distribution of interest $F_Y(Y|do(a))$ is identifiable when the copula $C(U_A, U_Y)$ is known so as to adjust for the non-causal back-door path between A and Y [5]. However, the copula, although uniquely exists [19], remains unknown and cannot be estimated from observational data as the noises are unobserved. Since the copula is unknown and unlearnable, it is inevitable to make assumptions about it to achieve causal effect identification. Specifically, the copula may be chosen from any of the vast families in the literature such as Archimedean, elliptical, or empirical copulas [19,40,41]. For instance, the Gaussian copula is one of the most studied elliptical copulas and it has been widely used in the fields of quantitative finance [42–44], hydrology research [45,46], logistics [47], astronomy [48], and others [19,40,41]. Recent works such as [49,50] have proposed sensitivity analysis with the Gaussian copula, but without the use of normalizing flows. In this work, we thus approximate the unknown copula $C(U_A, U_Y)$ with the Gaussian copula. That is,

$$C(U_A, U_Y) \approx \Phi_\rho(\Phi^{-1}(U_A), \Phi^{-1}(U_Y))$$

which yields

$$F_{Z_A,Z_Y}(Z_A, Z_Y) = C(\Phi(Z_A), \Phi(Z_Y)) \approx \Phi_\rho(Z_A, Z_Y)$$

where $\rho \in [-1, +1]$ is the Pearson's correlation between Z_A and Z_Y. See Fig. 1. As we will see, the latter expression fits really well into normalizing flows for modeling the SCM in Equation (4). It is also worth noticing that the Gaussian copula parameter ρ approximately denotes the confounding strength ρ_S, since $\rho = 2\sin(\pi\rho_S/6)$ [51,52].

3.3. ACE estimation with normalizing flows

Since F_{ε_A}, F_{ε_Y}, and Φ are strictly increasing functions, we can rewrite Equation (4) as

$$A := T_A^{-1}(Z_A)\ ,\ \ Y := T_{Y|A}^{-1}(Z_Y)\ ,\ \ (Z_A, Z_Y) \sim \Phi_\rho(Z_A, Z_Y) \tag{8}$$

where T_A and $T_{Y|A}$ are strictly monotonic, and thus invertible, transformations of the arbitrarily distributed observed random vector (A, Y) into the normally distributed random vector (Z_A, Z_Y). Such transformations are aptly called normalizing flow, and they are typically modeled as deep neural networks [53–55]. Specifically, we use unconstrained monotonic neural networks for the transformations [56,17], and the graphical conditioner neural network for conditioning the transformation of Y on its parent A [16,17]. Such a normalizing flow is able to universally model any arbitrary data distribution [57]. We henceforth refer to it as ρ-GNF.

As any normalizing flow, our ρ-GNF is trained by maximizing the log-likelihood of a observational dataset for a fixed ρ value. More specifically, let $X = (A, Y)$, $Z = (Z_A, Z_Y)$, and $T(Z_A, Z_Y; \theta) = (T_A(Z_A; \theta_A), T_{Y|A}(Z_Y; \theta_Y))$ where we explicitly represent the parameters of the deep neural networks as $\theta = (\theta_A, \theta_Y)$. Let $\{X^\ell\}_{\ell=1}^N$ denote an observational dataset. Then, the log-likelihood can be expressed as follows by a change of variables [53–55]:

$$\mathcal{LL}(\theta) = \sum_{\ell=1}^{N} \log\left(f_X(X^\ell; \theta)\right)$$

where

$$f_X(X^\ell; \theta) = f_Z\left(T^{-1}(X^\ell; \theta)\right) \cdot \left|\det\left(J_{T^{-1}(X^\ell;\theta)}(X^\ell)\right)\right|.$$

Recall that T is invertible by construction of normalizing flows. For the same reason, the determinant of the Jacobian $\det\left(J_{T^{-1}(X^\ell;\theta)} \times (X^\ell)\right)$ can be computed efficiently [53–55]. Under our Gaussian copula assumption, $f_Z(Z^\ell)$ is simply a bivariate normal density function. Therefore, our ρ-GNF can be trained efficiently.

Our main objective in this work is to estimate the average causal effect (ACE) as a function of ρ, which can be expressed as

$$ACE_\rho = E[Y|do(A := 1)] - E[Y|do(A := 0)] = E[Y_1] - E[Y_0]$$

where Y_a denotes the potential outcome under the intervention $A := a$. In particular, we use Monte-Carlo estimation by drawing samples from the interventional distribution $F_Y(Y|do(a))$ after having trained the ρ-GNF for a fixed ρ value. More concretely, we follow the following three steps:

1. $Z_Y^\ell = T_{Y|A^\ell}(Y^\ell; \theta_Y)$ for $\ell = 1, \ldots, N$.
2. $Y_a^\ell = T_{Y|a}^{-1}(Z_Y^\ell; \theta_Y)$ for $\ell = 1, \ldots, N$.
3. $ACE_\rho = E[Y_1] - E[Y_0] \approx \frac{\sum_{\ell=1}^N Y_1^\ell}{N} - \frac{\sum_{\ell=1}^N Y_0^\ell}{N}$.

The first step recovers the unobserved noise values for the observations in the training data, the second step computes the potential outcomes for the recovered noises, and the third step produces the Monte-Carlo estimates. The first step is possible due to the invertibility of normalizing flows. The first step can also be replaced by drawing N samples from $\Phi_\rho(Z_A, Z_Y)$.

3.4. Sensitivity analysis with ρ-GNF

As mentioned before, our main aim in this work is to estimate the ACE as a function of ρ, so as to determine how sensitivity the ACE is to different degrees of confounding. We have shown how to do it with the ρ-GNF. The result can be summarized in a sensitivity plot (ACE against ρ) that we aptly refer to as the ρ_{curve}. Note that, unlike in other sensitivity analysis methods, our sensitivity parameter ρ is conveniently bounded and interpretable. We illustrate this with examples in Section 5.

One of the most popular sensitivity analysis methods is based on the so-called E-value, which is defined as the minimum strength of association on the risk ratio scale that an unmeasured confounder would need to have with both treatment and outcome to fully explain away the observed treatment–outcome association, conditional on the measured covariates [20]. A large E-value thus implies that considerable unmeasured confounding would be needed to nullify the causal effect. Likewise, a small E-value implies that little unmeasured confounding would be needed to nullify the causal effect. Similar to the E-value, we propose the ρ_{value} which represents the Gaussian copula parameter value that explains away the observed association between treatment and outcome. In other words, setting $\rho = \rho_{value}$ results in $ACE_\rho = 0$. Let $\rho_{S_{Obs}}$ denote the Spearman correlation between the treatment A and the outcome Y in the observational data at hand. When A has no causal effect on Y, we have that $\rho_S(Z_A, Z_Y) = \rho_{S_{Obs}}$ by the scale-invariance property of Spearman correlation to the strictly increasing transformations T_A^{-1} and $T_{Y|A}^{-1} = T_Y^{-1}$. Therefore, $\rho_{value} = 2sin\left(\pi\rho_{S_{Obs}}/6\right)$.

Moreover, the ρ_{curve} and ρ_{value} may help the analyst to determine the sign of the ACE (i.e., whether the treatment is harmful or beneficial), which is arguably the most important part of sensitivity analysis. Specifically, suppose the analyst hypothesizes a measure of confounding in the interval $[\rho_{min}, \rho_{max}]$. Thus, the ρ_{curve} enables her to bound the ACE to the narrower interval $[ACE_{\rho_{max}}, ACE_{\rho_{min}}]$, which may in turn help her to determine the sign of the ACE: If $\rho_{min} > \rho_{value}$ (or $\rho_{max} < \rho_{value}$) then she may conclude that the ACE is negative (or positive). We illustrate this with examples in Section 5.

3.5. On the Gaussian copula assumption

We close this section with a discussion on the Gaussian copula assumption. In principle, any copula may be assumed in place of the Gaussian copula in Equations (5)-(7). However, the Gaussian copula assumption makes it possible to seamlessly integrate these equations into normalizing flows and thus produce our ρ-GNF, which may be seen as a generalization of the ordinary normalizing flow [53–55]. Specifically, the ordinary normalizing flow corresponds to ρ-GNF with $\rho = 0$. The fact that Z_A and Z_Y are still normally distributed makes the ρ-GNF retain one of the most salient features of the ordinary normalizing flow, namely the efficient computation of the log-likelihood of the training dataset which enables computationally efficient training of the ρ-GNF. Furthermore, the Gaussian copula assumption enables efficient sampling of (Z_A, Z_Y) for Monte-Carlo estimation of the ACE, thus enabling computationally efficient causal inference. Finally, the Gaussian copula assumption provides a single bounded and interpretable sensitivity parameter $\rho \in [-1, +1]$ for sensitivity analysis that can be used to control/model/adjust the back-door non-causal association under which the ACE is identifiable. Thus, enabling simple and efficient sensitivity analysis. Our subsequent experiments and results show that the Gaussian copula assumption works well empirically.

4. Bayesian sensitivity analysis with ρ-GNF

In this section, we extend our method for sensitivity analysis by approaching it from a Bayesian perspective. Defining a prior distribution over the sensitivity parameter (i.e., the ρ parameter of the Gaussian copula) makes it possible to derive the posterior distribution over the ACE, which makes it possible to calculate credible intervals.

Let Q be the causal quantity the analyst is interested in. In this work, Q is the ACE but our method actually applies to any causal quantity computable from the ρ-GNF (e.g., the expected potential outcome under some treatment). Then, calculating Q with the ρ-GNF will result in different values of Q for different values of ρ. Consequently, we can define Q as a function h of ρ such that $Q = h(\rho)$. However, we do not have any analytic expression of the function h, but we can evaluate it for a discrete set of points $\rho \in \mathcal{P} = \{\rho_1, \rho_2, \ldots, \rho_n\}$ by training the ρ-GNF for those values of ρ and evaluating Q. Next, we present two approaches for estimating h from the evaluation points $\mathcal{P}$, to be used to produce the posterior distribution of Q.

Continuous function approximation The first approach that we present is to make use of a change of variable to express the posterior distribution over Q as

$$f_Q(Q = q) = f_\rho\left(h^{-1}(q)\right) \cdot \left|\frac{dh^{-1}(q)}{dq}\right| \tag{9}$$

where $f_\rho(\rho)$ is a prior distribution over ρ, and then approximate h as a continuous function $\tilde{h}$ by using some method to interpolate the values $[-1, 1] \setminus \mathcal{P}$ for which h is not evaluated. For example, $\tilde{h}$ could be a piecewise linear function with n degrees of freedom such that $\tilde{h}(\rho) = h(\rho)$ for $\rho \in \mathcal{P}$.

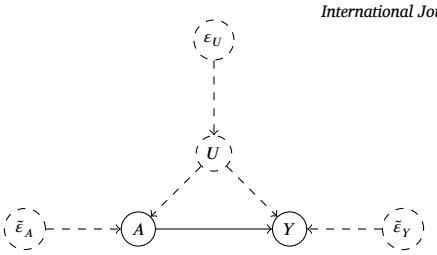


Fig. 2. Causal graph of the SCM in Proposition 1.

Unfortunately, this approach comes with several disadvantages. First, it is only possible if the inverse of h exists. This might not always be true in reality. Secondly, even if the inverse of h exists, it may be too badly approximated. For instance, let us suppose that $h(\rho) = \rho$. Due to the stochastic training process of the ρ-GNF, we may get some bias for our evaluation points so that $\tilde{h}(\rho) = h(\rho) + \epsilon$ where ϵ represents the bias. Then, for the derivative in point $\rho_0 \in \mathcal{P}$, we have that

$$\frac{d\tilde{h}(\rho_0)}{d\rho} \approx \lim_{\rho_0 - \tilde{\rho} \to 0} \frac{\tilde{h}(\rho_0) - \tilde{h}(\tilde{\rho})}{\rho_0 - \tilde{\rho}} = \lim_{\rho_0 - \tilde{\rho} \to 0} \frac{\rho_0 + \epsilon_0 - \tilde{\rho} - \tilde{\epsilon}}{\rho_0 - \tilde{\rho}} \in \{-\infty, 1, +\infty\} \ .$$

Observe that the derivative is $-\infty$ or $+\infty$ almost surely. It is easy to verify that the derivative of the inverse of $\tilde{h}$ suffers from the same limitation. This is problematic when plugged into Equation (9).

Discrete function approximation We present another approach for estimating h that involves representing it as a discrete function. Let the discrete approximation of h be denoted $\tilde{h} : \mathcal{P} \to \mathcal{Q} \subset \mathbb{R}$ with inverse $\tilde{h}^{-1} : \mathcal{Q} \to \mathcal{P}$. We furthermore redefine ρ as a discrete random variable taking values in $\mathcal{P}$. Without loss of generality, we assume that $\rho_i < \rho_{i+1}$ for $i = 1, \dots, n-1$ and define

$$P(\rho_i) = F_\rho\left(\frac{1}{2} \cdot (\rho_i + \rho_{i+1})\right) - F_\rho\left(\frac{1}{2} \cdot (\rho_i + \rho_{i-1})\right) \text{ for } i = 2, \dots, n-1$$
$$P(\rho_1) = F_\rho\left(\frac{1}{2} \cdot (\rho_1 + \rho_2)\right)$$
$$P(\rho_n) = 1 - F_\rho\left(\frac{1}{2} \cdot (\rho_n + \rho_{n-1})\right)$$

where F_ρ is the CDF of ρ corresponding to the prior distribution over ρ. Then, we can derive $P(Q = q)$ for $q \in \mathcal{Q}$ as

$$P(Q = q) = P\left(\rho \in \tilde{h}^{-1}(q)\right) = \sum_{r \in \tilde{h}^{-1}(q)} P(\rho = r). \tag{10}$$

Note that because we assume that ρ is discrete, Q becomes a discrete random variable as well. We can transform the discrete distribution of Q back to a continuous distribution by using kernel smoothing [58]:

$$f_Q(Q = q) = \sum_{\tilde{q} \in \mathcal{Q}} K(q - \tilde{q}) \cdot P(Q = \tilde{q}) \tag{11}$$

where K is a kernel function.

4.1. On the choice of the prior distribution of ρ

In this section, we discuss how a suitable prior distribution for ρ can be chosen to reflect expert knowledge about the data generating process. There are two main reasons for the noise variables ε_A and ε_Y to be dependent in the SCM in Equation (1) and Fig. 1: Selection bias and hidden confounding [5]. Selection bias occurs when conditioning on some common effect of the noise variables. Hidden confounding occurs when the noise variables have some common cause. As the latter is the most commonly studied scenario in the literature [5], we focus on it hereinafter. We start by discussing what different values of ρ imply about the hidden confounder. Based on these observations, we then make suggestions on how to do the reverse, i.e., how to derive a prior function over ρ that matches the existing expert knowledge about the hidden confounder.

Proposition 1. *The SCM in Equation (8) and depicted in Fig. 1 can be rewritten as an observationally and interventionally equivalent SCM with a hidden confounder as depicted in Fig. 2 by expressing it in the form of the following SCM:*

$$U := \varepsilon_U$$
$$A := \tilde{t}_A\left(U, \tilde{\varepsilon}_A\right) = T_A^{-1}\left(\frac{1}{\sqrt{\gamma^2 + \delta^2}}\left(\gamma U + \delta \tilde{\varepsilon}_A\right)\right)$$
$$Y := \tilde{t}_Y\left(U, A, \tilde{\varepsilon}_Y\right) = T_{Y|A}^{-1}\left(A, \lambda\rho U + \tau\sqrt{1 - \rho^2}\tilde{\varepsilon}_Y\right)$$

$$\varepsilon_U, \tilde{\varepsilon}_A, \tilde{\varepsilon}_Y \overset{iid}{\sim} \mathcal{N}(0,1)$$
$$\gamma \in \mathbb{R} \setminus \{0\}$$
$$\delta \in \left[-\sqrt{\frac{(1-\rho^2)\gamma^2}{\rho^2}},\ \sqrt{\frac{(1-\rho^2)\gamma^2}{\rho^2}} \right]$$
$$\lambda = \frac{\sqrt{\gamma^2+\delta^2}}{\gamma}$$
$$\tau = \sqrt{\frac{1}{1-\rho^2} - \frac{\gamma^2+\delta^2}{\gamma^2} \cdot \frac{\rho^2}{1-\rho^2}}\ .$$

The proof of the proposition can be found in Appendix A. The main idea is to rewrite the Gaussian noise random vector (Z_A, Z_Y) as a function of three independent Gaussian noise random variables $\tilde{\varepsilon}_A, \tilde{\varepsilon}_Y$ and ε_U.

Let the influence of the random variable U on another variable X be defined as $\frac{\partial X}{\partial U}$.

Theorem 1. *The influence of U on A and the influence of U on Y in the SCM in Proposition 1 have the same sign if and only if $\rho > 0$. The specific sign depends on the choice of γ.*

The proof of the theorem can be found in Appendix B.

Corollary 1. *For the SCM in Proposition 1, we cannot determine the sign of the effect of U on A or of the effect of U on Y, because the sign changes with the value of γ and any value of $\gamma \in \mathbb{R} \setminus 0$ yields the same observational and interventional distribution.*

Based on the above statements, we now make suggestions for a suitable ρ prior based on the believed influence of the hidden confounder on the treatment and outcome variables. We know that $\rho \in [-1, +1]$, and thus we need to choose a distribution whose PDF is zero elsewhere. Suitable prior functions are therefore for example a scaled and shifted beta distribution $f_\rho(\rho) = \frac{1}{2} f_{\text{Beta}}\left(\frac{1}{2}(\rho+1)\right)$ or a truncated normal distribution that is truncated between -1 and $+1$. The choice of ρ depends on the believed influence of the hidden confounder on the treatment and outcome variables. If one has no prior beliefs about it, a uniform prior between -1 and +1 is the natural choice. This equates to the beta distribution with parameters $\alpha = \beta = 1$ as shown in Fig. 3a.

From Theorem 1 we know that, believing that there exists a hidden confounder that has either a positive or a negative influence on both the treatment and the outcome, one should choose a prior distribution that puts more weight on positive values. This could for example be a truncated normal distribution with positive mean or a beta distribution with $\alpha > \beta$ as shown in Figs. 3b or 3c. If one believes that the influence of the confounder on the treatment has the opposite sign compared to the influence of the confounder on the outcome, one should choose a prior that puts more weight on negative values of ρ, like a beta distribution with $\alpha < \beta$ as shown in Fig. 3d.

In the case where one suspects that there is no confounding, one can choose a prior with mean zero (e.g., beta distribution with $1 < \alpha = \beta$ as shown in Fig. 3e). In the opposite case, where one suspects confounding but does not have any prior beliefs about the sign of the influence of the confounder on treatment and outcome, a prior that puts weight on values close to -1 and +1 should be chosen (e.g., beta distribution with $0 < \alpha = \beta < 1$ shown in Fig. 3f). We illustrate these scenarios with examples in the next section.

5. Experiments

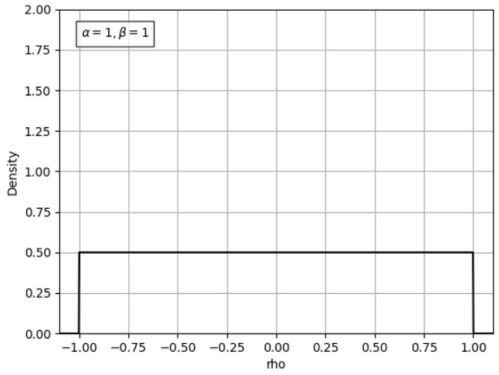
We present results for three different experimental settings: Simulated data with continuous outcome in Section 5.1, simulated data with binary outcome in Section 5.2, and real-world data with categorical outcome in Section 5.3. We present experiments for the ρ-GNF as well as for the Bayesian ρ-GNF.

Our ρ-GNF is implemented in PyTorch[1] by adapting the baseline code of the unconstrained monotonic neural networks [56][2] and the graphical normalizing flows [16].[3] As normalizing flows are developed for continuous variables, we use the Gaussian dequantization trick from the causal graphical normalizing flow to model discrete variables into ρ-GNF [17].

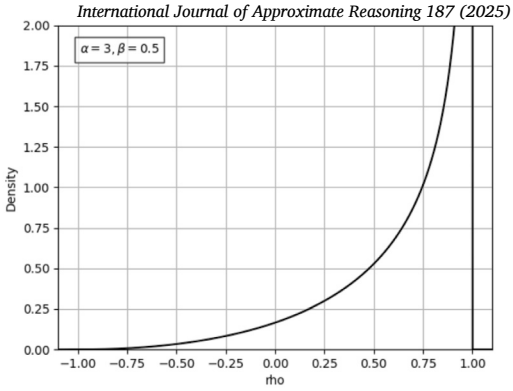
The ρ_{curve} for a given observational dataset is obtained by training ρ-GNFs for $\rho = -0.99, -0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8, 0.99$. We estimate the respective ACE_ρ as described in Section 3.3. Our empirical ACE bounds are obtained as the infimum and supremum of the ρ_{curve}, i.e., $\inf[ACE_\rho] \leq ACE_{true} \leq \sup[ACE_\rho]$. We also identify the ρ_{value} that explains away the causal association.

All experiments with the Bayesian ρ-GNF approach are run with values $\rho = -0.99, -0.95, -0.9, -0.85, -0.8, -0.75, \ldots, 0.75, 0.8, 0.85, 0.9, 0.95, 0.99$. In the experiments, Q is either the ACE or an expected potential outcome. Equations (10) and (11) are used to

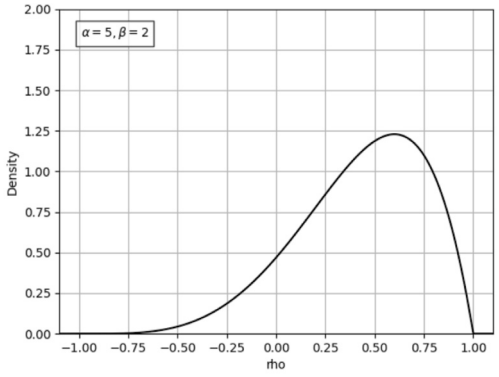
[1] The ρ-GNF code is available at https://github.com/sobalgi/rhoGNF.
[2] Code at https://github.com/AWehenkel/Graphical-Normalizing-Flows.
[3] Code at https://github.com/AWehenkel/UMNN.



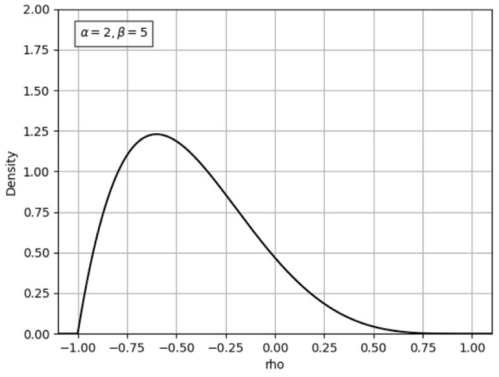
(a) Uniform prior.



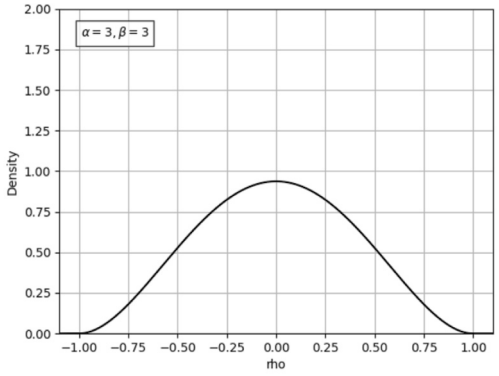
(b) Influence has same sign.



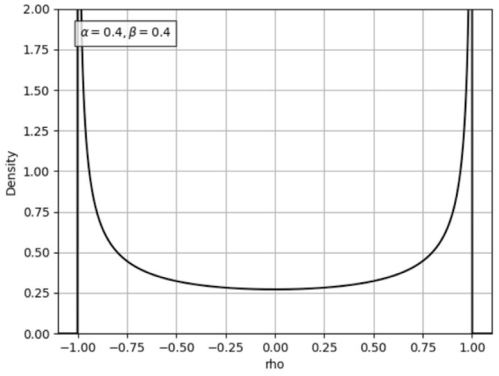
(c) Influence has same sign.



(d) Influence has opposite sign.



(e) No hidden confounding.



(f) Hidden confounding with unknown influence.

Fig. 3. Scaled and shifted beta distribution with different parameters α and β suitable for using as ρ prior. The subcaptions describe the beliefs about hidden confounding that are encoded in the distributions. The term "influence" in the subcaption refers to the influence of the hidden confounder on the treatment and outcome variables.

estimate the distribution of the ACE and the expected potential outcome. As kernel function, we use a Gaussian kernel with variance $\frac{1}{16}\text{Var}(Q)$ whenever plotting the density of Q.

5.1. Experiments with continuous outcome

In our first set of simulated experiments, we consider the SCM with continuous treatment A and outcome Y in the work [59]. This is a well-studied SCM from economics and econometrics, and is defined as follows.

Table 1
Six observationally and/or interventionally non-equivalent SCMs with different mixtures of total observed association ($\rho_{P_{Obs}}$), non-causal association (ρ_{true}), causal association (ACE_{true}), and ρ_{value}. Note that $ACE_{true}=\alpha$.

$SCM_{\alpha,\beta,\delta}$	α	β	δ	$\rho_{P_{Obs}}$	ρ_{true}	ACE_{true}	ρ_{value}
SCM_1	0.2	-0.6	0.72	-0.55	-0.71	0.2	-0.55
SCM_2	0.0	-0.4	0.52	-0.55	-0.55	0.0	-0.55
SCM_3	-0.2	-0.2	0.40	-0.55	-0.32	-0.2	-0.55
SCM_4	0.2	0.2	0.40	0.55	0.32	0.2	0.55
SCM_5	0.0	0.4	0.52	0.55	0.55	0.0	0.55
SCM_6	-0.2	0.6	0.72	0.55	0.71	-0.2	0.55

$$A:=\varepsilon_A \ , \quad Y:=\alpha A+\varepsilon_Y \ , \begin{pmatrix}\varepsilon_A\\ \varepsilon_Y\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \beta\\ \beta & \delta\end{pmatrix}\right) \ . \tag{12}$$

For given α, β and δ values, we have that the total Pearson's correlation between A and Y is $\rho_{P_{Obs}}=\frac{\sigma_{A,Y}}{\sigma_A\sigma_Y}$, where $\sigma_A^2=1$ and $\sigma_Y^2=\alpha^2+\delta+2\alpha\beta$ and $\sigma_{A,Y}=\alpha+\beta$. The Pearson's correlation due to the non-causal path is $\rho_{true}=\frac{\sigma_{\varepsilon_A,\varepsilon_Y}}{\sigma_{\varepsilon_A}\sigma_{\varepsilon_Y}}=\beta/\delta$, while the causal one is $ACE_{true}=\alpha$. These values are used for verification purposes. See Table 1 for the α, β and δ values considered in our experiments.

Fig. 4a shows the scatter plot of each of the six observational datasets obtained by sampling the six SCMs in Table 1. Each dataset contains 50,000 observations. As expected, the SCMs that are observationally equivalent present similar observational data distributions, even though the SCMS are not interventionally equivalent. Likewise, Fig. 4b shows that the datasets corresponding to observationally equivalent SCMs result in similar ρ_{curve} plots. From this figure, we can note that $\rho=\rho_{value}$ implies $ACE_\rho=0$ and $\rho=\rho_{true}$ implies $ACE_\rho=ACE_{true}$, as can be verified with the help of Table 1. All these observations confirm the accuracy of our ρ-GNF for causal inference.

The SCM in Equation (12) is a simple example involving a continuous outcome and linear relationships, similar to the ones studied in the works [25,26]. However, our ρ-GNF can learn arbitrary non-linear transformations, and thus our sensitivity analysis method also applies to non-linear SCMs, as we demonstrate with further experiments involving binary and categorical outcomes in the next sections. Before that, we show the results of analyzing SCM_1 in Table 1 with our Bayesian ρ-GNF.

Specifically, we estimate the posterior distribution of the ACE for two different ρ priors. Adopting a uniform distribution over ρ, Fig. 5a shows that the ACE appears to be negative. On the other hand, when we adopt a truncated normal distribution that is centered around the true ρ value (i.e., ρ_{true} in Table 1), we can see in Fig. 5b that the distribution over the ACE shifts towards the true ACE value (i.e., ACE_{true} in Table 1).

In practice, researchers might also be interested in estimating the probability for the treatment A having a positive effect on the outcome Y. Because our Bayesian ρ-GNF provides us with the full posterior distribution of the ACE, we can calculate $P(ACE>0)$. For the uniform prior we get $P(ACE>0)=0.21$ while for the truncated normal prior we get $P(ACE>0)=0.61$.

5.2. Experiments with binary outcome

For our second set of simulated experiments, we consider a setting used in previous works on sensitivity analysis [22–24]. Namely, a SCM with binary treatment A, binary outcome Y, and binary confounder U. The SCM is randomly parameterized by sampling $\{P(U), P(A|U), P(Y|A,U)\}$ uniformly from the interval $[0,1]$. We do so 20 times to produce 20 SCMs that are sampled to produce 20 observational datasets over the random vector (A,Y) by discarding the values for U. Each dataset contains 50,000 observations. For evaluation purposes, we compute the true ACE (i.e., ACE_{true}) for the SCMs sampled by adjusting for U [5]. From the datasets sampled, we can only compute the assumption free (AF) bounds of ACE_{true} as follows, since U is not observed:

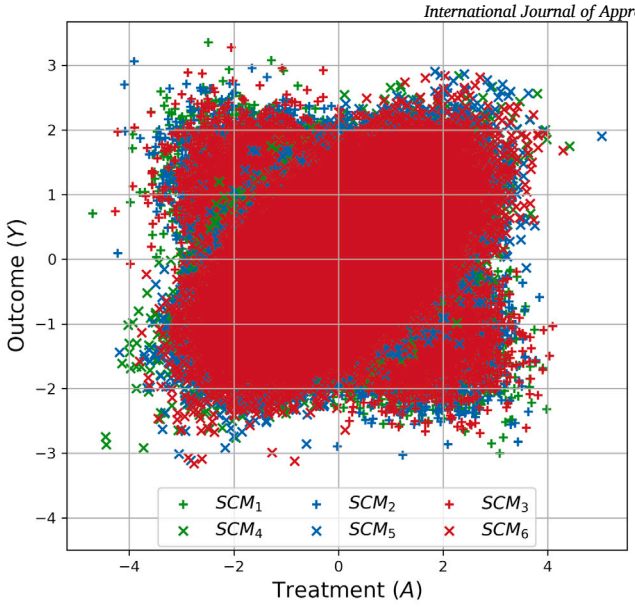
$$AF_{lower}=q_1p_1-q_0p_0-p_1 \leq ACE_{true} \leq AF_{upper}=q_1p_1-q_0p_0+p_0 \tag{13}$$

where $p_a=P(A=a)$ and $q_a=P(Y=1|A=a)$ are estimated from the data [21,8].

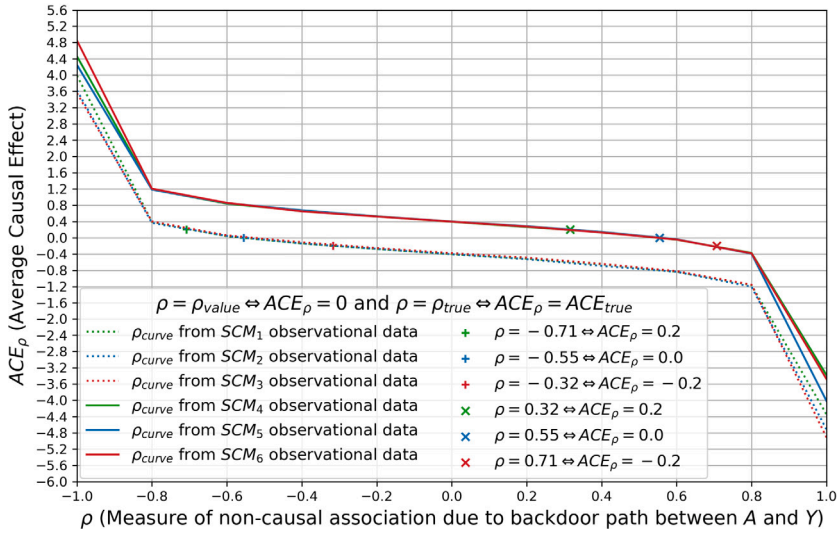
Fig. 6 shows eight ρ_{curve} plots that are representative of the results obtained. Note that for each ρ_{curve}, the ACE bounds obtained as the infimum and supremum of the curve (i.e., $\inf[ACE_\rho]$ and $\sup[ACE_\rho]$) include ACE_{true}. As expected due to the Gaussian copula assumption, these bounds are narrower than the AF bounds. These observations together with the ones made in the previous section confirm that our ρ-GNF is accurate for both continuous and binary outcomes. This generality distinguishes our method from the existing sensitivity analysis methods. The next section presents experiments with a categorical outcome and real-world data.

5.3. Experiments with real-world datasets

In this section, we present experiments with two real-world datasets. First, we use our ρ-GNF approach to analyze the impact of the International Monetary Fund on child poverty [60–63]. Afterwards, we present a sensitivity analysis for potential outcomes with the classical Blau and Duncan social mobility dataset [64].



(a) Scatter plots of the observational datasets sampled from the six SCMs in Table 1.



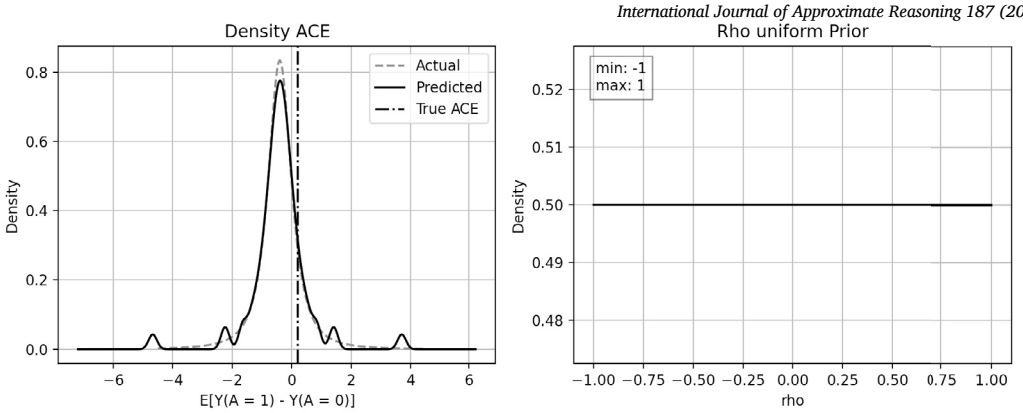
(b) Six ρ_{curve} plots from the observational datasets sampled from the six SCMs in Table 1.

Fig. 4. Observationally equivalent SCMs, i.e., $\{SCM_1, SCM_2, SCM_3\}$ and $\{SCM_4, SCM_5, SCM_6\}$, show similar scatter plots and ρ_{curve} plots. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

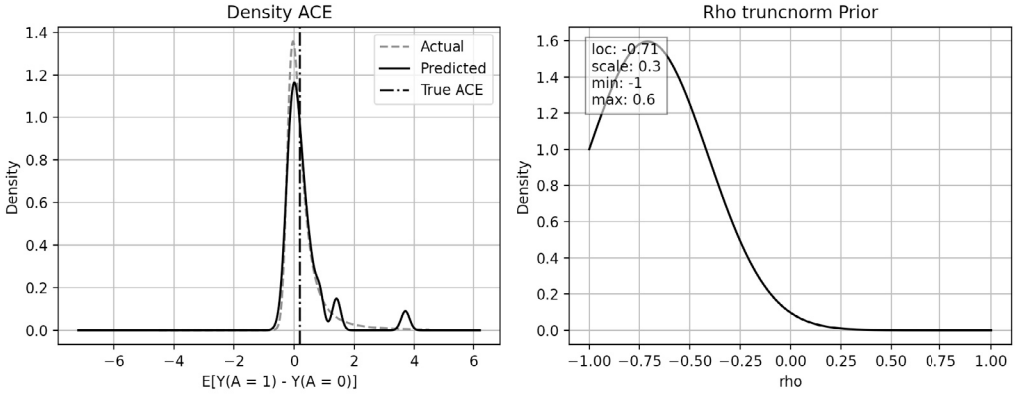
5.3.1. *International monetary fund*

The International Monetary Fund (IMF) is an international organization that aims to promote global macroeconomic stability through a series of programs. However, the impact of these programs on children is a subject of debate [61,65,66]. In this section, we study the impact of the programs (treatment A) on child poverty (outcome Y). We consider the IMF child poverty dataset used previously in the works [60–63]. It contains 1,941,734 observations each corresponding to a child under the age of 18 residing in 67 countries from the Global-South region, which includes the least developed countries.[4] For each child, the dataset records whether she receives the treatment (i.e., she lives in a country adopting the IMF program) as well as her degree of poverty. The degree of

[4] Due to the sensitive nature and the accompanying ethical considerations, the dataset is not publicly available but it can be requested upon from the original authors.



(a) Uniform prior.



(b) Truncated normal prior.

Fig. 5. Shown on the left is the posterior distribution of the ACE for the ρ prior shown on the right. The distribution labeled "actual" is the theoretical convergence limit for $n \to \infty$ of a universal function approximator like the ρ-GNF when trained on n data points assuming the given ρ prior. The curve labeled "predicted" shows the distribution estimated with an actual ρ-GNF trained on $n = 100,000$ data samples.

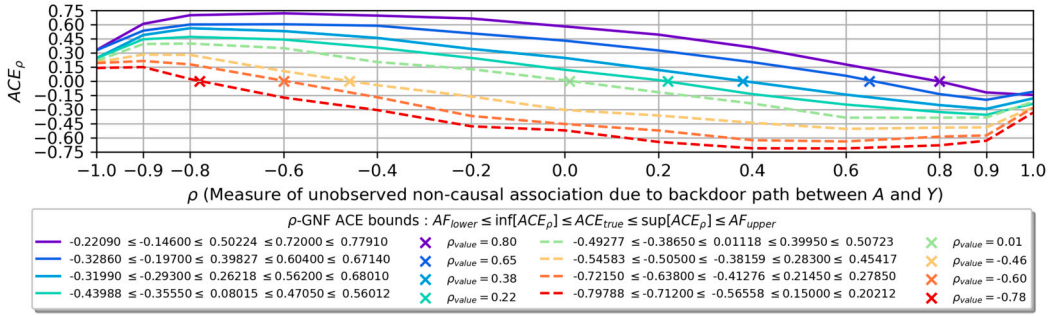


Fig. 6. Eight ρ_{curve} plots corresponding to eight of the observational datasets in the experiments with binary outcome, and their respective ρ_{value} and bounds.

poverty ranges from 0 (no poverty) to 7 (severe poverty). It is calculated as the sum of seven binary indicators of poverty, representing access to education, health services, information, sanitation, shelter, food and water.

It is most likely that the IMF dataset is subject to unmeasured confounding, and thus the true ACE is not identifiable from the data [5]. However, we can bound it by computing the AF bounds in Equation (13) for each of the seven binary indicators of poverty, and then summing them. This results in $AF_{lower} = -3.34$ and $AF_{upper} = 3.56$.

Fig. 7 shows the ρ_{curve} produced by our sensitivity analysis method. As expected due to the Gaussian copula assumption, the bounds $\inf[ACE_\rho] = -1.73$ and $\sup[ACE_\rho] = 2.11$ obtained from the curve are narrower than the AF bounds. The fact that the

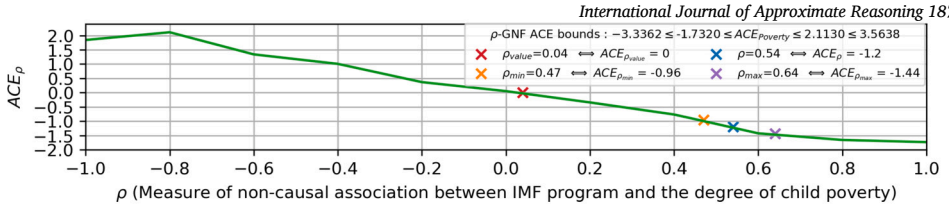


Fig. 7. ρ_{curve} for the degree of the child poverty, and its ρ_{value} and bounds.

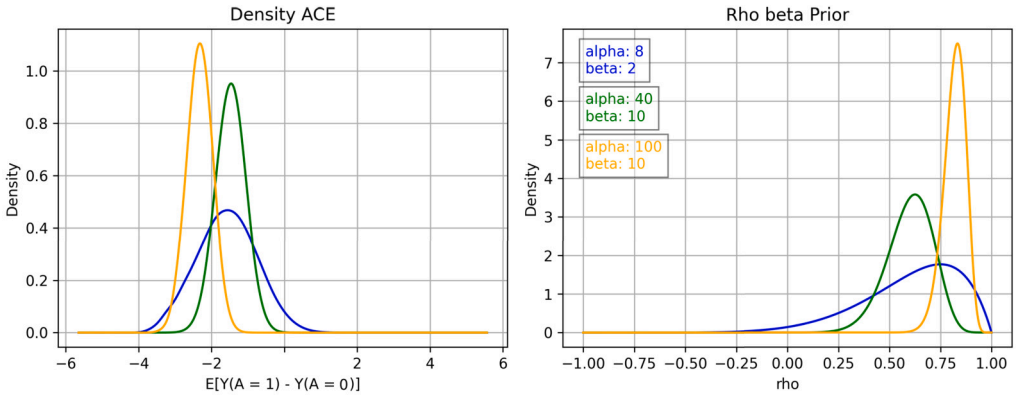


Fig. 8. Estimated posterior distributions of the ACE assuming a scaled and shifted beta distribution as ρ prior with different values for α and β.

ρ_{value} is as small as 0.04 indicates that the unmeasured confounding and the causal effect cancel each other, i.e., they have equal magnitude but opposite signs. In other words, either there is no unmeasured confounding and no causal effect of the IMF program on child poverty, or both exist and have the same strength but opposite signs. Previous works rule out the first explanation [60–63]. Moreover, the second explanation is not unreasonable as we elaborate next. A country with severe socioeconomic turmoil (i.e., large ε_Y) causing severe child poverty is associated with higher incentives to apply for the IMF program (i.e., large ε_A) to overcome the socioeconomic turmoil. Similarly, a country with less socioeconomic turmoil (i.e., small ε_Y) has less incentives to apply for the IMF program (i.e., small ε_A). Therefore, it is not unreasonable to assume that ε_A and ε_Y are positively associated, which correspond to assuming a positive ρ value in our sensitivity analysis framework. In that case, the true ACE would be negative as can be seen in Fig. 7. For instance, if the domain experts were to believe that $\rho \in [0.47, 0.64]$, then $ACE_{true} \in [-1.44, -0.96]$ which indicates that the IMF program is beneficial for reducing child poverty. This real-world example illustrates how our ρ-GNF helps the analyst to obtain ACE bounds that are more informative than the AF bounds. Our Bayesian ρ-GNF can provide the analyst with a even more detailed picture, as we show below.

Specifically, since ε_A and ε_Y are arguably positively associated, we choose a scaled and shifted beta distribution with three different parameter combinations of α and β that all have positive mean as ρ prior. Fig. 8 shows the prior distributions and the resulting posterior distributions of the ACE. Comparing the blue and green posterior distributions that have the same mean but different variance, we can conclude from the plot that the more certain one is in the value of ρ, the smaller the variance of the ACE. Comparing the green to the yellow posterior distributions, we can also conclude that shifting the mean of the prior towards larger values of ρ implies lower ACE values. Our Bayesian analysis affirms the beneficial nature of the IMF program in the reduction of child poverty, as for all three prior distributions with positive mean, most of the posterior density lies on negative ACE values. For example, the blue prior with $\alpha = 8$ and $\beta = 2$ results in a 95% credible interval for the ACE of $[-3.177, -0.057]$.

5.3.2. *Blau and Duncan social mobility dataset*

The work [64] is considered a pioneering study of intergenerational social mobility in the U.S. The authors used linear path analysis (i.e., linear SCMs) to analyze data from the 1962 "Occupational Changes in a Generation" survey of 20,000 men aged 20–64. They examined how family background, education, and early career achievements influenced occupational outcomes. The study assumed the causal graph depicted in Fig. 9, except for the bidirected edge between U and Y.

The focus of the original study was on Y, i.e., the occupational status of the son's job in 1962 when the survey was conducted. The original causal graph did not contain any hidden confounding involving Y. However, as argued in the work [67], it is not unreasonable to assume that the son's educational attainment U and Y are confounded by motivation or grit, which is not measured in the dataset.[5] This is represented by adding the bidirected edge between U and Y to the original causal graph in Fig. 9. Both U and Y are categorical random variables. The former has categories 0-8, and the latter 0-96.

[5] Note that the hidden confounder is typically denoted as U in the literature, while we stick to the notation of the original paper where U is the son's educational attainment.

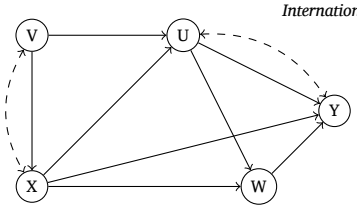


Fig. 9. The causal graph corresponding to the Blau and Duncan dataset, except for the bidirected edge between U and Y. Bidirectional edges indicate correlation between the corresponding unmodeled ϵ nodes. The nodes in the graph represent the father's educational attainment (V), father's occupational status (X), son's educational attainment (U), the occupational status of the son's first job (W), and the occupational status of the son's job in 1962 (Y), i.e., the year the data were collected.

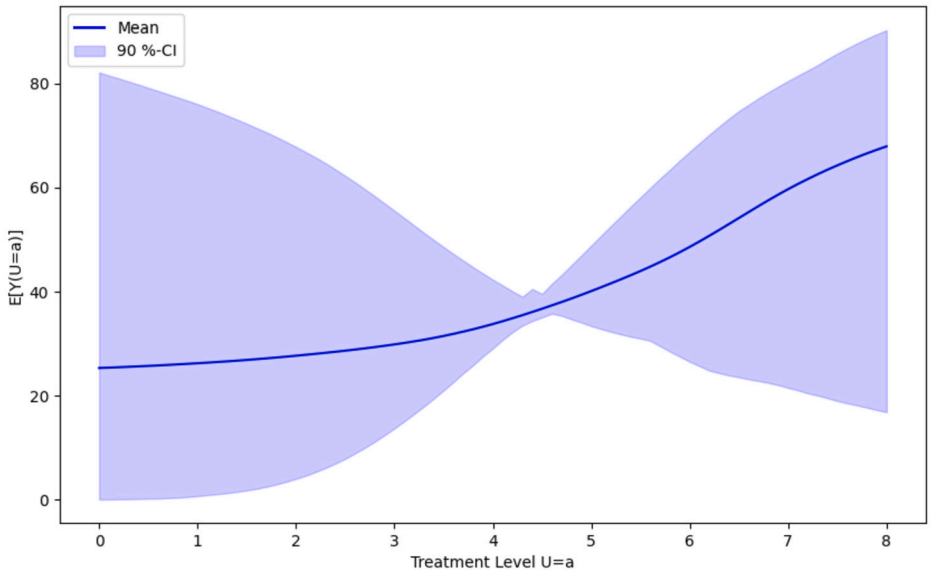


Fig. 10. Posterior mean of the potential outcome of Y and 90% credible interval assuming a uniform ρ prior between -1 and 1.

We run our Bayesian ρ-GNF framework to estimate the expected potential outcome of Y for different treatment levels of U, as a function of the strength of their unmeasured confounder. As discussed above, this confounder may well be motivation or grit. However, it may be motivation just for educational achievements, or just for work achievements, or both, or none. In other words, we are uncertain about the sign of the influence of this confounder on U and Y, and thus we adopt a uniform ρ prior in the interval $[-1, 1]$ following the recommendations in Section 4.1. We compute the posterior distribution of the potential outcome of Y for different treatment levels of U. Fig. 10 plots the posterior mean and 90% credible intervals. We can conclude that U seems to have an effect on Y, and this effect seems to be non-linear. Moreover, the effect seems to be rather insensitive towards confounding at treatment levels around 4-5 but quite sensitive at higher and lower treatment levels. This is partially explained by the fact that the more extreme the treatment level is the fewer the observations in the dataset. This example illustrates that our Bayesian ρ-GNF framework is an informative tool for sensitivity analysis with categorical treatment and/or outcomes. This together with our previous experiments with continuous and binary outcomes demonstrate the generality of our framework, which is a feature that distinguishes it from the existing sensitivity analysis methods.

6. Conclusion

We proposed ρ-GNF, a novel approach for sensitivity analysis to unobserved confounding that is based on copula-based normalizing flows. Our approach contains a bounded and interpretable sensitivity parameter ρ representing the unobserved non-causal association between the observed treatment and outcome due to confounding. Under the Gaussian copula assumption, we showed that ρ-GNF enabled us to estimate the causal effect as a function of ρ in the form of the ρ_{curve}. The ρ_{curve} enabled us to identify the ρ_{value}, i.e., the confounding strength needed to nullify the causal effect. This is related to the E-value in the literature. The ρ_{curve} also enabled us to identify empirical bounds of the causal effect that are narrower than the assumption free bounds in the literature.

We also proposed a Bayesian version of ρ-GNF by defining a prior distribution over the sensitivity parameter ρ, which allowed us to calculate the posterior distribution over any quantity of interest that can be deducted from the ρ-GNF. This enabled us to derive credible intervals for the causal effect. We also discussed how to choose a suitable prior based on expert knowledge about the hidden confounder.

Finally, we illustrated the benefits of our sensitivity analysis method with simulated and real-world data under different settings, e.g., with continuous, binary and categorical outcomes. This generality distinguishes our method from the methods in the literature.

It is worth recalling from the discussion in Section 3.5 that, although the Gaussian copula fits our framework particularly well, any other copula may be used. As a matter of fact, we recommend using several copulas and integrate the conclusions drawn from all of them, in order to increase robustness. Likewise, we recommend integrating too the results of other sensitivity analysis methods in the literature, which are based on different assumptions than ours. This process is sometimes called triangulation. We do not pursue this further in this paper because it is a known practice [6].

It is also worth commenting on the scalability of the framework presented in this paper to large datasets. Our experiments in Section 5.3.1 with a dataset of almost 2 million observations were run in an ordinary computer, since no particularly deep neural networks were required. Specifically, the neural networks used had three fully connected hidden layers of between 5 and 20 units each. This proves that our framework scales well.

Finally, it is worth reminding the reader of the following caveat against over-interpreting causal conclusions, including the ones in this work. Sensitivity analysis rarely rules out the possibility that the causal effect is null, i.e., that the observed association between exposure and outcome is not solely due to confounding. Instead, sensitivity analysis informs the user of the confounding strength required to nullify the causal effect, and it is up to the user to decide whether such strength is likely or not in the domain at hand. In our Bayesian framework, we similarly let the user specify a prior distribution over the confounding strength and produce a posterior distribution over the causal effect to decide if the null causal effect is likely or not. Therefore, any causal claim is based on the assumptions made by the sensitivity analysis method considered and the final judgment of the user about the confounding strength required to nullify the causal effect. As mentioned above, one way to increase the strength of causal claims is by triangulation [6], i.e., by integrating the results of several sensitivity analysis methods based on potentially different assumptions.

CRediT authorship contribution statement

Sourabh Balgi: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marc Braun:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Jose M. Peña:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Adel Daoud:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jose M. Peña, Adel Daoud, Sourabh Balgi reports financial support was provided by Swedish Research Council (ref. 2019-00245). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Proof of Proposition 1

First, note that δ needs to be bound between $\pm\sqrt{\frac{(1-\rho^2)\gamma^2}{\rho^2}}$ as otherwise $\frac{1}{1-\rho^2} - \frac{\gamma^2+\delta^2}{\gamma^2} \cdot \frac{\rho^2}{1-\rho^2}$ would be negative, and consequently τ would not be a real number. Likewise, γ needs to be non-zero as otherwise δ or $\tilde{\varepsilon}_A$ are $0/0$. Furthermore, observe that when

$$\begin{pmatrix} \frac{1}{\sqrt{\gamma^2+\delta^2}}\left(\gamma U + \delta\tilde{\varepsilon}_A\right) \\ \lambda\rho U + \tau\sqrt{1-\rho^2}\tilde{\varepsilon}_Y \end{pmatrix} = \begin{pmatrix} Z_A \\ Z_Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

then the SCMs in Equation (1) and in the proposition are observationally and interventionally equivalent. We know that because $\tilde{\varepsilon}_A, \tilde{\varepsilon}_Y, \varepsilon_U \overset{iid}{\sim} \mathcal{N}(0,1)$, any linear combination of them is also normally distributed. It therefore suffices to show that the mean vector is $(0,0)^T$, the variances are 1, and the covariance is ρ.

First, note that

$$\begin{aligned} E\left[\frac{1}{\sqrt{\gamma^2+\delta^2}}\left(\gamma U + \delta\tilde{\varepsilon}_A\right)\right] &= \frac{1}{\sqrt{\gamma^2+\delta^2}}\left(\gamma E[U] + \delta E\left[\tilde{\varepsilon}_A\right]\right) \\ &= \frac{1}{\sqrt{\gamma^2+\delta^2}}\left(\gamma\cdot 0 + \delta\cdot 0\right) \\ &= 0\ . \end{aligned}$$

Second, note that

$$\begin{aligned} E\left[\lambda\rho U + \tau\sqrt{1-\rho^2}\tilde{\varepsilon}_Y\right] &= \lambda\rho E[U] + \tau\sqrt{1-\rho^2}E\left[\tilde{\varepsilon}_Y\right] \\ &= \lambda\rho\cdot 0 + \tau\sqrt{1-\rho^2}\cdot 0 \end{aligned}$$

$$= 0 \ .$$

Third, note that

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{\sqrt{\gamma^2+\delta^2}}\left(\gamma U + \delta\tilde{\varepsilon}_A\right)\right) &= \frac{1}{\gamma^2+\delta^2}\left(\gamma^2\mathrm{Var}(U) + \delta^2\mathrm{Var}(\tilde{\varepsilon}_A)\right)\\ &= \frac{\gamma^2+\delta^2}{\gamma^2+\delta^2}\\ &= 1 \ .\end{aligned}$$

Fourth, note that

$$\begin{aligned}\mathrm{Var}(\lambda\rho U + \tau\sqrt{1-\rho^2}\tilde{\varepsilon}_Y) &= \lambda^2\rho^2\mathrm{Var}(U) + \tau^2(1-\rho^2)\mathrm{Var}(\tilde{\varepsilon}_Y)\\ &= \lambda^2\rho^2 + \tau^2(1-\rho^2)\\ &= \frac{\gamma^2+\delta^2}{\gamma^2}\rho^2 + \left(\frac{1}{1-\rho^2} - \frac{\gamma^2+\delta^2}{\gamma^2}\cdot\frac{\rho^2}{1-\rho^2}\right)\cdot(1-\rho^2)\\ &= \frac{(\gamma^2+\delta^2)\rho^2}{\gamma^2} + \left(1 - \frac{(\gamma^2+\delta^2)\rho^2}{\gamma^2}\right)\\ &= 1 \ .\end{aligned}$$

Finally, note that

$$\begin{aligned}\mathrm{Cov}\left(\frac{1}{\sqrt{\gamma^2+\delta^2}}\left(\gamma U + \delta\tilde{\varepsilon}_A\right), \lambda\rho U + \tau\sqrt{1-\rho^2}\tilde{\varepsilon}_Y\right) &= \frac{\gamma\cdot\rho\cdot\lambda}{\sqrt{\gamma^2+\delta^2}}\mathrm{Cov}(U,U)\\ &\quad + \frac{\gamma\cdot\sqrt{1-\rho^2}\cdot\tau}{\sqrt{\gamma^2+\delta^2}}\mathrm{Cov}(U,\tilde{\varepsilon}_Y)\\ &\quad + \frac{\lambda\cdot\delta\cdot\rho}{\sqrt{\gamma^2+\delta^2}}\mathrm{Cov}(\tilde{\varepsilon}_A,U)\\ &\quad + \frac{\delta\cdot\tau\cdot\sqrt{1-\rho^2}}{\sqrt{\gamma^2+\delta^2}}\mathrm{Cov}(\tilde{\varepsilon}_A,\tilde{\varepsilon}_Y)\\ &= \frac{\gamma}{\sqrt{\gamma^2+\delta^2}}\cdot\lambda\cdot\rho\\ &= \frac{\gamma}{\sqrt{\gamma^2+\delta^2}}\cdot\frac{\sqrt{\gamma^2+\delta^2}}{\gamma}\cdot\rho\\ &= \rho \ .\end{aligned}$$

Appendix B. Proof of Theorem 1

Recall that $T_A^{-1}(Z_A)$ is strictly increasing in Z_A by construction. Then, we have that

$$\begin{aligned}\mathrm{sgn}\left(\frac{\partial\tilde{t}_A}{\partial U}\right) &= \mathrm{sgn}\left(\frac{\partial Z_A}{\partial U}\cdot\frac{dT_A^{-1}(Z_A)}{dZ_A}\right)\\ &= \mathrm{sgn}\left(\frac{\gamma}{\sqrt{\gamma^2+\delta^2}}\right)\cdot\mathrm{sgn}\left(\frac{dT_A^{-1}(Z_A)}{dZ_A}\right)\\ &= \mathrm{sgn}\left(\frac{\gamma}{\sqrt{\gamma^2+\delta^2}}\right)\cdot 1\\ &= \mathrm{sgn}(\gamma) \ .\end{aligned}$$

Recall that $T_{Y|A}^{-1}(Z_Y)$ is strictly increasing in Z_Y by construction. Then, we have that

$$\begin{aligned}\mathrm{sgn}\left(\frac{\partial\tilde{t}_Y}{\partial U}\right) &= \mathrm{sgn}\left(\frac{\partial Z_Y}{\partial U}\cdot\frac{dT_{Y|A}^{-1}(Z_Y)}{dZ_Y}\right)\\ &= \mathrm{sgn}(\lambda\rho)\cdot\mathrm{sgn}\left(\frac{dT_{Y|A}^{-1}(Z_Y)}{dZ_Y}\right)\end{aligned}$$

$$= \operatorname{sgn}\left(\rho \frac{\sqrt{\gamma^2+\delta^2}}{\gamma}\right) \cdot 1$$
$$= \operatorname{sgn}(\rho) \cdot \operatorname{sgn}(\gamma) \quad .$$

Since $\gamma \neq 0$ by definition, we then have that

$$\rho > 0 \implies \operatorname{sgn}\left(\frac{\partial \tilde{t}_A}{\partial U}\right) = \operatorname{sgn}\left(\frac{\partial \tilde{t}_Y}{\partial U}\right)$$
$$\rho \leq 0 \implies \operatorname{sgn}\left(\frac{\partial \tilde{t}_A}{\partial U}\right) \neq \operatorname{sgn}\left(\frac{\partial \tilde{t}_Y}{\partial U}\right) \quad .$$

Data availability

Data will be made available on request.

References

[1] S. Wright, Correlation and causation, J. Agric. Res. 20 (1921) 557–585.
[2] R.A. Fisher, Design of experiments, Br. Med. J. 1 (1936) 554.
[3] D.R. Cox, Planning of Experiments, Wiley, 1958.
[4] G.W. Imbens, D.B. Rubin, Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction, Cambridge University Press, 2015.
[5] J. Pearl, Causality: Models, Reasoning and Inference, Cambridge University Press, 2009.
[6] M.A. Hernán, J.M. Robins, Causal Inference: What If, Chapman & Hall/CRC, 2020.
[7] J.J. Schlesselman, Assessing effects of confounding variables, Am. J. Epidemiol. 108 (1) (1978) 3–8.
[8] C.F. Manski, Nonparametric bounds on treatment effects, Am. Econ. Rev. 80 (2) (1990) 319–323.
[9] G.W. Imbens, Sensitivity to exogeneity assumptions in program evaluation, Am. Econ. Rev. 93 (2) (2003) 126–132.
[10] B.A. Brumback, M.A. Hernán, S.J. Haneuse, J.M. Robins, Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures, Stat. Med. 23 (5) (2004) 749–767.
[11] T.J. VanderWeele, O.A. Arah, Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders, Epidemiology (2011) 42–52.
[12] J. Cornfield, W. Haenszel, E.C. Hammond, A.M. Lilienfeld, M.B. Shimkin, E.L. Wynder, Smoking and lung cancer: recent evidence and a discussion of some questions, J. Natl. Cancer Inst. 22 (1) (1959) 173–203.
[13] W.G. Cochran, D.B. Rubin, Controlling bias in observational studies: a review, Sankhya, Ser. A (1973) 417–446.
[14] K.J. Rothman, S. Greenland, T.L. Lash, Modern Epidemiology, Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008.
[15] T.L. Lash, M.P. Fox, A.K. Fink, et al., Applying Quantitative Bias Analysis to Epidemiologic Data, Springer, 2009.
[16] A. Wehenkel, G. Louppe, Graphical normalizing flows, in: International Conference on Artificial Intelligence and Statistics (AISTATS), 2021, pp. 37–45.
[17] S. Balgi, J.M. Peña, A. Daoud, Personalized public policy analysis in social sciences using causal-graphical normalizing flows, in: AAAI Conference on Artificial Intelligence (AAAI), 2022, pp. 11810–11818.
[18] A. Javaloy, P. Sánchez-Martín, I. Valera, Causal normalizing flows: from theory to practice, in: Neural Information Processing Systems (NeurIPS), 2023, pp. 58833–58864.
[19] R.B. Nelsen, An Introduction to Copulas, Springer, 2007.
[20] T.J. VanderWeele, P. Ding, Sensitivity analysis in observational research: introducing the E-value, Ann. Intern. Med. 167 (4) (2017) 268–274.
[21] J.M. Robins, The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies, in: Health Service Research Methodology: A Focus on AIDS, 1989, pp. 113–159.
[22] A. Sjölander, A note on a sensitivity analysis for unmeasured confounding, and the related E-value, J. Causal Inference 8 (1) (2020) 229–248.
[23] A. Sjölander, O. Hössjer, Novel bounds for causal effects based on sensitivity parameters on the risk difference scale, J. Causal Inference 9 (1) (2021) 190–210.
[24] J.M. Peña, Simple yet sharp sensitivity analysis for unmeasured confounding, J. Causal Inference 10 (1) (2022) 1–17.
[25] C. Cinelli, D. Kumor, B. Chen, J. Pearl, E. Bareinboim, Sensitivity analysis of linear structural causal models, in: International Conference on Machine Learning (ICML), 2019, pp. 1252–1261.
[26] C. Cinelli, C. Hazlett, Making sense of sensitivity: extending omitted variable bias, J. R. Stat. Soc., Ser. B Stat. Methodol. 82 (1) (2020) 39–67.
[27] S. Balgi, J.M. Peña, A. Daoud, ρ-GNF: a copula-based sensitivity analysis to unobserved confounding using normalizing flows, in: International Conference on Probabilistic Graphical Models (PGM), 2024, pp. 20–37.
[28] P. Ding, T.J. VanderWeele, Sensitivity analysis without assumptions, Epidemiology 27 (3) (2016) 368.
[29] V. Veitch, A. Zaveri, Sense and sensitivity analysis: simple post-hoc analysis of bias due to unobserved confounding, in: Neural Information Processing Systems (NeurIPS), 2020, pp. 10999–11009.
[30] A. Sjölander, S. Greenland, Are E-values too optimistic or too pessimistic? Both and neither!, Int. J. Epidemiol. (2022).
[31] M. Ilse, P. Forré, M. Welling, J.M. Mooij, Combining interventional and observational data using causal reductions, arXiv preprint, arXiv:2103.04786, 2021.
[32] J. Ioannidis, Y. Tan, M. Blum, Limitations and misinterpretations of E-values for sensitivity analyses of observational studies, Ann. Intern. Med. 170 (2) (2019) 108–111.
[33] E.J.T. Tchetgen, I. Shpitser, Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis, Ann. Stat. 40 (3) (2012) 1816.
[34] A. Lindmark, X. de Luna, M. Eriksson, Sensitivity analysis for unobserved confounding of direct and indirect effects using uncertainty intervals, Stat. Med. 37 (10) (2018) 1744–1762.
[35] J. Peters, D. Janzing, B. Schölkopf, Elements of Causal Inference: Foundations and Learning Algorithms, The MIT Press, 2017.
[36] J.E. Angus, The probability integral transform and related results, SIAM Rev. 36 (4) (1994) 652–654.
[37] J.M. Mooij, J. Peters, D. Janzing, J. Zscheischler, B. Schölkopf, Distinguishing cause from effect using observational data: methods and benchmarks, J. Mach. Learn. Res. 17 (1) (2016) 1103–1204.
[38] C. Spearman, The proof and measurement of association between two things, Int. J. Epidemiol. 39 (5) (2010) 1137–1150.
[39] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1/2) (1938) 81–93.
[40] G. Salvadori, C. De Michele, N.T. Kottegoda, R. Rosso, Extremes in Nature: An Approach Using Copulas, Springer, 2007.
[41] F. Durante, C. Sempi, Principles of Copula Theory, CRC Press, 2016.

[42] U. Cherubini, E. Luciano, W. Vecchiato, Copula Methods in Finance, John Wiley & Sons, 2004.
[43] F. Salmon, Recipe for disaster: the formula that killed Wall Street, Wired Magazine 17 (3) (2009).
[44] D. MacKenzie, T. Spears, The formula that killed wall street: the Gaussian copula and modelling practices in investment banking, Soc. Stud. Sci. 44 (3) (2014) 393–417.
[45] B. Renard, M. Lang, Use of a Gaussian copula for multivariate extreme value analysis: some case studies in hydrology, Adv. Water Resour. 30 (4) (2007) 897–912.
[46] L. Zhang, V.P. Singh, Copulas and Their Applications in Water Resources Engineering, Cambridge University Press, 2019.
[47] P. Kumar, Copula Functions and Applications in Engineering, in: Logistics, Supply Chain and Financial Predictive Analytics, Springer, 2019, pp. 195–209.
[48] T.T. Takeuchi, Constructing a bivariate distribution function with given marginals and correlation: application to the galaxy luminosity function, Mon. Not. R. Astron. Soc. 406 (3) (2010) 1830–1840.
[49] J. Zheng, A. D'Amour, A. Franks, Copula-based sensitivity analysis for multi-treatment causal inference with unobserved confounding, arXiv preprint, arXiv: 2102.09412, 2021.
[50] J. Zheng, J. Wu, A. D'Amour, A. Franks, Sensitivity to unobserved confounding in studies with factor-structured outcomes, arXiv preprint, arXiv:2208.06552, 2022.
[51] W.H. Kruskal, Ordinal measures of association, J. Am. Stat. Assoc. 53 (284) (1958) 814–861.
[52] C. Meyer, The bivariate normal copula, Commun. Stat., Theory Methods 42 (13) (2013) 2402–2422.
[53] G. Papamakarios, I. Murray, T. Pavlakou, Masked autoregressive flow for density estimation, in: Neural Information Processing Systems (NeurIPS), 2017, pp. 2338–2347.
[54] G. Papamakarios, E. Nalisnick, D.J. Rezende, S. Mohamed, B. Lakshminarayanan, Normalizing flows for probabilistic modeling and inference, J. Mach. Learn. Res. 22 (57) (2021) 1–64.
[55] I. Kobyzev, S. Prince, M. Brubaker, Normalizing flows: an introduction and review of current methods, IEEE Trans. Pattern Anal. Mach. Intell. 43 (11) (2021) 3964–3979.
[56] A. Wehenkel, G. Louppe, Unconstrained monotonic neural networks, in: Neural Information Processing Systems (NeurIPS), 2019, pp. 1545–1555.
[57] C. Huang, D. Krueger, A. Lacoste, A.C. Courville, Neural autoregressive flows, in: International Conference on Machine Learning (ICML), 2018, pp. 2083–2092.
[58] M.P. Wand, M.C. Jones, Kernel Smoothing, CRC Press, 1994.
[59] K.D. Hoover, Causality in Economics and Econometrics, SSRN eLibrary, 2006.
[60] A. Daoud, F.D. Johansson, The impact of austerity on children: uncovering effect heterogeneity by political, economic, and family factors in low- and middle-income countries, Soc. Sci. Res. 118 (2024) 102973.
[61] A. Daoud, E. Nosrati, B. Reinsberg, A.E. Kentikelenis, T.H. Stubbs, L.P. King, Impact of international monetary fund programs on child health, Proc. Natl. Acad. Sci. USA 114 (25) (2017) 6492–6497.
[62] B. Halleröd, B. Rothstein, A. Daoud, S. Nandy, Bad governance and poor children: a comparative analysis of government efficiency and severe child deprivation in 68 low-and middle-income countries, World Dev. 48 (2013) 19–31.
[63] S. Balgi, J.M. Peña, A. Daoud, Counterfactually-equivalent structural causal modelling using causal graphical normalizing flows, in: International Conference on Probabilistic Graphical Models (PGM), 2024, pp. 164–181.
[64] P.M. Blau, Duncan, The American Occupational Structure, John Wiley and Sons, 1967.
[65] A. Daoud, B. Reinsberg, Structural adjustment, state capacity and child health: evidence from IMF programmes, Epidemiology 48 (2) (2018) 445–454.
[66] A. Daoud, B. Reinsberg, A.E. Kentikelenis, T.H. Stubbs, L.P. King, The International Monetary Fund's Interventions in Food and Agriculture: An Analysis of Loans and Conditions, vol. 83, Food Policy, 2019, pp. 204–218.
[67] S. Balgi, A. Daoud, J.M. Peña, G. Wodtke, J. Zhou, Deep learning with DAGs, Sociol. Methods Res. (2025).