On the Identifiability and Estimation of Causal Location-Scale Noise Models

Alexander Immer [1 2]   **Christoph Schultheiss** [3]   **Julia E Vogt** [1]   **Bernhard Schölkopf** [1 2]
Peter Bühlmann [3]   **Alexander Marx** [1 4]

Abstract

We study the class of location-scale or het-eroscedastic noise models (LSNMs), in which the effect Y can be written as a function of the cause X and a noise source N independent of X, which may be scaled by a positive function g over the cause, i.e., $Y = f(X) + g(X)N$. Despite the generality of the model class, we show the causal direction is identifiable up to some pathological cases. To empirically validate these theoretical findings, we propose two estimators for LSNMs: an estimator based on (non-linear) feature maps, and one based on neural networks. Both model the conditional distribution of Y given X as a Gaussian parameterized by its natural parameters. When the feature maps are correctly specified, we prove that our estimator is jointly concave, and a consistent estimator for the cause-effect identification task. Although the the neural network does not inherit those guarantees, it can fit functions of arbitrary complexity, and reaches state-of-the-art performance across benchmarks.

1. Introduction

Distinguishing cause from effect, given only observational data, is a fundamental problem in many natural sciences such as medicine or biology, and lies at the core of causal discovery. Without any prior knowledge, distinguishing whether X causes Y ($X \rightarrow Y$), or whether Y causes X ($Y \rightarrow X$) is unattainable (Pearl, 2000). When assuming a properly restricted structural causal model (SCM), however, the true graph that generated the data is identifiable (Peters et al., 2011). In its most general form, a structural causal

[1]Department of Computer Science, ETH Zurich, Switzerland [2]Max Planck Institute for Intelligent Systems, Tübingen, Germany [3]Seminar for Statistics, ETH Zurich, Switzerland [4]AI Center, ETH Zurich, Switzerland. Correspondence to: Alexander Immer <alexander.immer@inf.ethz.ch>, Alexander Marx <alexander.marx@inf.ethz.ch>.
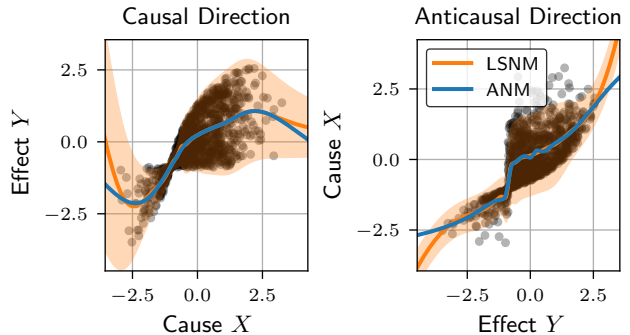
Figure 1. Location-scale and additive noise model fits on MNU pair 55. Both models show a similar mean regression but the LSNM can model the scale of the variance. Therefore, the LSNM correctly identifies the causal direction with a log-likelihood difference of ≈ 0.155 while the ANM identifies the wrong direction with difference ≈ -0.001.

model expresses the effect Y as a function of the cause X and an independent noise term N, that is, $Y = f(X, N)$.

There exists a vast literature that derived assumptions under which restricted SCMs are identifiable. That is, there exists no backward model which fulfills the same modeling assumptions as the assumed forward model. For simple linear additive noise models (ANMs), for example, it has been shown that if the data are non-Gaussian, then there exists no backward model, such that the cause can be modeled as a linear function of the effect and an additive independent noise term (Shimizu et al., 2006). Instead, for all such backward models, the noise will depend on the effect. Apart from linear models, identifiability results have been established for non-linear ANMs (Hoyer et al., 2009; Bühlmann et al., 2014), where $Y = f(X) + N$ and post non-linear (PNL) noise models (Zhang & Hyvärinen, 2009; Zhang et al., 2015), where $Y = f_2(f_1(X) + N)$. Besides identifiability, it has also been shown that consistent estimators for ANMs exist (Kpotufe et al., 2014).

Here, we focus on location-scale noise models (LSNMs) or heteroscedastic noise models,[1] where the effect Y is ex-

---

[1]Heteroscedastic noise refers to the setting where the variance of the noise is non-constant and *depends* on the value of X. To emphasize that we model N as an independent source scaled by a function over X, we use the notion of location-scale noise.

pressed as $Y = f(X) + g(X)N$. LSNMs generalize the classical ANM setting by allowing the noise source N to be scaled by a positive function $g(x)$. Naturally, if $g(x) = 1$, an LSNM simplifies to an ANM. In Fig. 1, we provide an example from a synthetic benchmark (Tagasovska et al., 2020), in which the ground truth model follows an LSNM. The ANM-based approach not only identifies the wrong direction, but also, the causal and anti-causal model have near identical log-likelihoods. Thus, the decision is merely a coin flip. The LSNM model, however, identifies the true direction with high confidence. Simply put, we can learn both $f(x)$ and $g(x)$ for the causal direction and can thus separate the independent noise source from X, whereas in the anti-causal direction we cannot find such a pair of functions. Hence, the estimated noise remains skewed.

The previous example assumes that we already know the answer to the following questions: Under which assumptions are LSNMs *identifiable*? How and under which assumptions can we *consistently estimate $f(x)$ and $g(x)$*? How can we combine these results to arrive at a score that allows us to *identify cause and effect from observational data*?

Contributions In this paper, we address each of the questions above. In Sec. 2, we formally define LSNMs and show that apart from some pathological cases—which are known for the special case of Gaussian noise (Khemakhem et al., 2021)—no backward model exists.[2]

In Sec. 3, we study the estimation of LSNMs under the Gaussian noise assumption, which relates to the problem of maximizing the conditional log-likelihood of Y given X via heteroscedastic regression. Typical estimators aim to learn $f(x)$ as the mean of the Gaussian and $g(x)$ as its variance, which leads to a non-concave likelihood function. We instead propose to relate $f(x)$ and $g(x)$ to the natural parametrization of a Gaussian, with corresponding jointly concave log-likelihood. Further, we show that we can consistently estimate LSNMs via (non-linear) feature maps if they are correctly specified. To allow for a more general class of functions, we propose a second estimator based on neural networks (NNs). Due to overparametrization, the NN-based estimator does not have guarantees. In practice, however, it reaches state-of-the-art performance, outperforming the variant based on feature maps.

To use our estimators for cause-effect inference, we propose LOCI (**lo**cation-scale **c**ausal **i**nference) in Sec. 4. Concretely, we propose a variant of LOCI via log-likelihood estimation, extending the approach of (Zhang & Hyvärinen, 2009) to LSNMs, and implement a variant of LOCI following the RESIT approach (regression with subsequent independence testing) (Peters et al., 2014). For both strategies, we show under which assumptions consistent estimation of cause and effect is attainable with our estimator based on feature maps.

We evaluate all variants of LOCI on standard cause-effect benchmark datasets and observe that LOCI achieves state-of-the-art overall performance and almost perfect accuracy on datasets which are in line with our assumptions—i.e., Gaussian ANMs and LSNMs.

For reproducibility, we make our code publicly available[3] and provide all proofs in the Supplementary Material.

2. Identifiability of LSNMs

In this section, we focus on the identifiability of location-scale noise models (LSNMs). A causal model is said to be identifiable under a set of structural constraints, if only the forward (causal) model is well specified and no backward model fulfilling these structural constraints exists.

To formally analyze this problem, we first need to define our assumed causal model.

Definition 1 (Location-Scale Noise Model) *Given two independent random variables X and N_Y. If the effect Y is generated by a location-scale noise model, we can express Y as an SCM of the form*

$$Y := f(X) + g(X)N_Y ,$$

where $f\colon \mathcal{X} \to \mathbb{R}$ and $g\colon \mathcal{X} \to \mathbb{R}_+$, i.e. g is strictly positive.

We provide an illustration of data generated by a location-scale noise model in Fig. 1. LSNMs simplify to ANMs when $g(X)$ is constant, and to multiplicative noise models when $f(X)$ is constant.

To prove identifiability of such a restricted SCM, it is common to derive an ordinary differential equation (ODE), which needs to be fulfilled such that a backward model exists, see e.g. Hoyer et al. (2009), or Zhang & Hyvärinen (2009). Intuitively, the solution space of such an ODE specifies all cases in which the model is non-identifiable, leaving all specifications which do not fulfill the ODE as identifiable. In the following theorem, we derive such a differential equation for LSNMs and discuss its implications.

Theorem 1 *Assume the data is such that a location-scale noise model can be fit in both directions, i.e.,*

$$Y = f(X) + g(X) N_Y, \quad X \perp\!\!\!\perp N_Y$$
$$X = h(Y) + k(Y) N_X, \quad Y \perp\!\!\!\perp N_X.$$

Let $\nu_1(\cdot)$ and $\nu_2(\cdot)$ be the twice differentiable log densities

[3]https://github.com/AlexImmer/loci

of Y and N_X respectively. For compact notation, define

$$\nu_{X|Y}(x|y) = \log\left(p_{X|Y}(x|y)\right)$$
$$= \log\left(p_{N_X}\left(\frac{x - h(y)}{k(y)}\right)/k(y)\right)$$
$$= \nu_2\left(\frac{x - h(y)}{k(y)}\right) - \log(k(y)) \quad and$$
$$G(x,y) = g(x)f'(x) + g'(x)[y - f(x)].$$

Assume that $f(\cdot)$, $g(\cdot)$, $h(\cdot)$, and $k(\cdot)$ are twice differentiable. Then, the data generating mechanism must fulfill the following PDE for all x, y with $G(x, y) \neq 0$.

$$0 = \nu_1''(y) + \frac{g'(x)}{G(x,y)}\nu_1'(y) + \frac{\partial^2}{\partial y^2}\nu_{X|Y}(x|y) + \quad (1)$$
$$\frac{g(x)}{G(x,y)}\frac{\partial^2}{\partial y \partial x}\nu_{X|Y}(x|y) + \frac{g'(x)}{G(x,y)}\frac{\partial}{\partial y}\nu_{X|Y}(x|y).$$

The equality derived in Theorem 1 is equivalent to the result concurrently provided by Strobl & Lasko (2022) up to fixing a sign error in the terms involving $g'(x)$. We derived this result independently using a different proof technique and additionally note that $p_{X|Y}(x|y)$ cannot be written as univariate function with argument $([x - h(y)]/k(y))$ if $Y \to X$ is an LSNM with non-constant $k(\cdot)$.

The conclusion of Strobl & Lasko (2022) is that if we have x_0 such that $G(x_0, y) \neq 0$ for all but countably many y, then knowing $\nu_{X|Y}(x_0|y)$, $G(x_0, y) \neq 0$, $g(x_0)$ and $g'(x_0)$ leads to $\nu_1(y)$ being constrained to a two dimensional affine space as (1) becomes an ODE. This is in analogy to the result of Hoyer et al. (2009) for ANMs. For this case, Zhang & Hyvärinen (2009) have refined the result and provide a list of all possible cases of unidentifiable models: only for specific choices of $f(\cdot)$ and $\nu_2(\cdot)$, one can find $\nu_1(\cdot)$ such that the model is invertible.

This conclusion carries over to the LSNM. Assume there exist different values x such that $G(x, y) \neq 0$ for all but countably many y. If $g(\cdot)$ is strictly positive and $f(\cdot)$ is injective, this applies to all $x \in \mathbb{R}$ except for at most countably many. Each such value leads to a different ODE in y when plugging it into Eq. (1). Only when the solution spaces of all ODEs overlap such that the same $\nu_1(\cdot)$ is found, which must also be valid log-density, the model can be invertible. This is not the case for generic combinations of $g(\cdot)$, $G(\cdot, \cdot)$ and $\nu_{X|Y}(\cdot|\cdot)$ but only for very specific exceptions. Thus, apart from some pathological cases, an LSNM cannot be invertible. A precise characterization of these cases as in (Zhang & Hyvärinen, 2009) for the post-nonlinear model, which involves the ANM as a special case, has not yet been found for LSNM to the best of our knowledge.

To provide a bit more intuition regarding the assumptions of Theorem 1, note that the results only apply to random

variables X with unbounded support. This is implied by requiring that the log-density of N_X has to be twice differentiable. For example, X could not follow a uniform distribution. This also implies that $g(\cdot)$ has to be a nonlinear (or constant) function since otherwise $g(\cdot)$ is negative for some attainable values of X and does not strictly map to $\mathbb{R}_+$ as required by our assumptions. Assuming that the noise variable is Gaussian, necessary conditions for the distributions of X and Y as well as the functions $f(\cdot)$, $g(\cdot)$, $h(\cdot)$, and $k(\cdot)$ can be found (Khemakhem et al., 2021). For completeness, we provide the corresponding result as Theorem 4 in Supplementary Material B.

3. Estimation of LSNMs

To separate the independent noise source $N_Y = \frac{Y - f(X)}{g(X)}$ from Y, we derive a consistent maximum likelihood estimator for the log-likelihood of Y given X and a parameter vector $\boldsymbol{\theta}$ that models $f(x)$ and $g(x)$. The typical parametrization then attempts to fit $f(x)$ as the mean of a Gaussian and $g(x)$ as the standard deviation, such as iterative feasible generalized least-squares (Harvey, 1976; Amemiya, 1985). However, this leads to non-concavity of the associated log-likelihood and therefore massively complicates consistency of the estimator. In particular, we would define $\theta_1(x) = \mu(x)$ and $\theta_2(x) = \sigma^2(x)$ and have the Gaussian log-likelihood $\log p(y|x, \boldsymbol{\theta}) = \log \mathcal{N}(y|\theta_1(x), \theta_2(x))$. Then, differentiating twice with respect to $\theta_2(x)$, we have

$$\frac{\partial^2 \log \mathcal{N}(y|\theta_1(x), \theta_2(x))}{\partial \theta_2(x)^2} = \frac{\theta_2(x) - 2(\theta_1(x) - y)^2}{2\theta_2(x)^3},$$

which is negative if $\theta_2(x) < 2(\theta_1(x) - y)^2$ since $\theta_2(x) > 0$ as it models the variance. Therefore, the Hessian can be indefinite and the log-likelihood cannot be guaranteed to be concave. With a non-concave estimator, deriving any kind of consistency claims is challenging (Wooldridge, 2015). Consistent estimation, however, is crucial to guarantee that an estimator can provably identify the causal direction.

To achieve consistent estimation and consequently provide an estimator that can identify the causal direction in the large-data limit, we choose to parameterize the Gaussian with its natural parameters, $\eta_1(x), \eta_2(x)$ with the inverse mapping $\mu(x) = -\frac{\eta_1(x)}{2\eta_2(x)}$ and $\sigma^2(x) = -\frac{1}{2\eta_2(x)}$. Due to properties of the natural parametrization of exponential families, it is clear that the Gaussian log-likelihood

$$\log p(y|x, \boldsymbol{\eta}(x)) = c + \boldsymbol{\eta}(x)^\top \begin{bmatrix} y \\ y^2 \end{bmatrix} + \frac{\eta_1(x)^2}{4\eta_2(x)} + \log\sqrt{-2\eta_2(x)}$$
(2)

with $c = -1/2\log(2\pi)$, is strictly concave in $\eta_1(x), \eta_2(x)$ (Brown, 1986) with constraint $\eta_2(x) < 0$. However, the parametrization of the natural parameters through learnable functions matters.

We are particularly interested in non-linear Gaussian LSNMs. We therefore assume to know the right feature maps giving rise to ground truth $\eta_1(x)$ and $\eta_2(x)$. With (non-linear) feature maps $\boldsymbol{\psi}(x) \in \mathbb{R}^D$, $\boldsymbol{\phi}(x) \in \mathbb{R}^D_+$, and parameters $\boldsymbol{w}_1 \in \mathbb{R}^D$, $\boldsymbol{w}_2 \in \mathbb{R}^D_+$, we have

$$\eta_1(x) = \boldsymbol{\psi}(x)^\mathsf{T}\boldsymbol{w}_1 \ \text{ and } \ \eta_2(x) = -\boldsymbol{\phi}(x)^\mathsf{T}\boldsymbol{w}_2 \quad (3)$$

with corresponding log-likelihood as $\log p(y|x, \boldsymbol{w})$. The second natural parameter, $\eta_2(x)$, is ensured to be negative due to positive feature map and parameters. The log-likelihood of this model is then jointly concave in $\boldsymbol{w}_1, \boldsymbol{w}_2$ due to the composition of a concave log-likelihood with a linear model and constraint $\boldsymbol{w}_2 > 0$.

Lemma 1 *With natural parameters modelled as $\eta_1(x) = \boldsymbol{\psi}(x)^\mathsf{T}\boldsymbol{w}_1$ and $\eta_2(x) = -\boldsymbol{\phi}(x)^\mathsf{T}\boldsymbol{w}_2$, the Gaussian log-likelihood function in Eq. (2) is jointly concave in $\boldsymbol{w}_1, \boldsymbol{w}_2$.*

The result relies primarily on the concavity of exponential family distributions in their natural form and has been used in a similar context for heteroscedastic Gaussian processes (Le et al., 2005). In contrast, the commonly used iterative feasible generalized least-squares (FGLS) method for heteroscedastic regression is formulated with a non-concave loss. For example, Cawley et al. (2004) use a similar formulation using feature maps but their model is not jointly concave because they do not use natural parameters.

To achieve consistent estimation, we need the log-likelihood to be strictly concave (to guarantee global identification) when given a certain number of samples $T > 0$. This condition can be formulated in terms of the Fisher information, which is the expectation of the negative log-likelihood Hessian under the predictive sample for each data point.

Assumption 1 *There exists $T_0 > 0$, s.t. for any $T \geq T_0$ the Fisher information matrix of the Gaussian log-likelihood in Eq. (2) parameterized as in Eq. (3), i.e.,*

$$I_T(\boldsymbol{w}) = -\frac{1}{T}\sum_{i=1}^{T}\mathbb{E}_{y\sim p(y|x_i;\boldsymbol{w})}\left[\nabla^2_{\boldsymbol{w}}\log p(y|x_i;\boldsymbol{w})\right] \quad (4)$$

is positive definite with respect to any $\boldsymbol{w} \in \mathbb{R}^D \times \mathbb{R}^D_+$.

In our case, the Fisher information and the Hessian coincide since we have a generalized linear model with exponential family likelihood (Sec. 9.2 in Martens, 2020). Therefore, the Fisher information is guaranteed to be at least positive semi-definite according to Lemma 1. In particular, each summand in Eq. (4) will have rank two. Thus, our assumption essentially requires that, upon observing enough data points, these summands span different sub-spaces of $\mathbb{R}^{2D}$ such that their sum is full rank. For sensible feature maps and non-degenerate distributions on x, we expect this assumption to hold. This is trivially the case for $D = 2$, which

could, for example, be a linear model with $\phi(x) = x$ and $\psi(x) = |x|$, because each summand is already full rank.

Theorem 2 *Let $\{(y_i, x_i)\}_{i=i,...,T}$ be an iid sample with conditional density $p(y|x_i; \boldsymbol{w}^*)$ as defined through Eqs. (2) and (3), with $\boldsymbol{w}^* \in \mathbb{R}^D \times \mathbb{R}^D_+$, and correctly specified (non-linear) feature maps $\boldsymbol{\psi}(x) \in \mathbb{R}^D$ and $\boldsymbol{\phi}(x) \in \mathbb{R}^D_+$. Further, there exists a $T_0 > 0$ for which Assumption 1 holds, and suppose the usual smoothness criteria hold:*

i) *Derivatives up to second order can be passed under the integral sign in $\int dP(y|x, \boldsymbol{w})$.*

ii) *Third derivatives $\nabla^3 \log p(y|x, \boldsymbol{w})$ are bounded by a function $M(y|x)$ on $\mathbb{R}^D \times \mathbb{R}^D_+$, s.t.*

$$\sup_{\theta \in \mathbb{R}^D \times \mathbb{R}^D_+}|\nabla^3 \log p(y|x, \boldsymbol{w})_{jkl}| \leq M(y|x)$$

for all j, k, l, and $\mathbb{E}[M(y|x)] < \infty$.

Then, for $T \geq T_0$ the maximum likelihood estimate $\hat{\boldsymbol{w}}_T$ is weakly consistent, i.e., for $T \to \infty$, $\hat{\boldsymbol{w}}_T \overset{p}{\to} \boldsymbol{w}^*$. Further, $\hat{\boldsymbol{w}}_T$ is asymptotically normal in a sense that $\sqrt{T}(\hat{\boldsymbol{w}}_T - \boldsymbol{w}^*) \overset{d}{\to} \mathcal{N}(0, I(\boldsymbol{w}^*)^{-1})$.

The key assumptions for this Theorem to hold are that the feature maps are correctly specified and that they fulfill Assumption 1, which we discussed above. The remaining assumptions ensure the smoothness of the log-likelihood and are needed to prove the theorem following the approach of Cramér (1946). There exists follow-up work, which shows that these smoothness criteria can be relaxed, e.g., Kulldorff (1957) and Wald (1949), however, we used the conditions provided by Cramér (1946) as they are more intuitive. Last, we can also achieve consistency (but not asymptotic normality) by substituting the smoothness criteria in Theorem 2 with the condition

$$\mathbb{E}[|\log p(y|x, \boldsymbol{w})|] < \infty \,,$$

as discussed in (Wooldridge, 2015, Ch. 15).

In Supplementary Material A, we describe an algorithm to estimate the model parameters. Due to the concavity in $\boldsymbol{w}$, the algorithm can make use of closed-form iterations and reliably converges to a global optimum. In general however, we expect our model to be misspecified and the underlying feature maps to be unknown. In this case, we use approximate feature maps for our linear model. For example, we use spline-based feature maps in our experiments (Eilers & Marx, 1996). In Supplementary Material E, we compare the performance of our proposed estimator to the commonly used iterative feasible generalized least squares (IFGLS) algorithm, which does not rely on a jointly concave log-likelihood formulation. We find that our estimator clearly

improves upon IFGLS. Especially, for small datasets, which are common in causal inference, and large numbers of feature maps it leads to significantly less overfitting.

In the case where we do not know the true feature maps, we can alternatively use neural networks to learn them. Parameterized by weights $\boldsymbol{w} \in \mathbb{R}^D$, we have neural network $\boldsymbol{f} : \mathbb{R} \times \mathbb{R}^D \to \mathbb{R}^2$ that maps from a scalar input to the two natural parameters of the heteroscedastic likelihood. In particular, we model

$$\eta_1(x) = f_1(x, \boldsymbol{w}) \text{ and } \eta_2(x) = -\tfrac{1}{2}\exp(f_2(x, \boldsymbol{w})),$$

which uses an additional exponential link function that ensures positivity while maintaining differentiability. For details, see Supplementary Material A.2.

4. Cause-Effect Inference

After studying identifiability and estimation of LSNMs, we now propose two strategies to instantiate our estimators: a likelihood-based estimator and an independence-based estimator. Both assume we can separate the independent noise N_Y from X, Y. Hence, we first show under which conditions maximizing Eq. (2) is equivalent to minimizing the dependence between estimated noise $\hat{N}$ and X.

One way to measure the dependence between two continuous random variables is mutual information

$$\mathrm{I}(X; Y) = \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy \ ,$$

which is zero iff X and Y are independent. Suppose there exists an invertible mapping from $(x, n_Y)^T$ to $(x, y)^T$, where the mapping function $\boldsymbol{f}$ is defined as an SCM with model class $\mathcal{F}$, and $\boldsymbol{f}$ can be parameterized by $\boldsymbol{\theta}$. For such an invertible mapping with corresponding Jacobian transformation matrix $\boldsymbol{J}_{X \to Y}$, mutual information (sample version) of X and N_Y can be expressed as (Zhang et al., 2015)

$$\mathrm{I}(\boldsymbol{x}, \boldsymbol{n}_Y | \boldsymbol{\theta}) = -\frac{1}{T} \sum_{i=1}^{T} \log \left(\frac{p(x_i)p(n_{Y,i}|\boldsymbol{\theta})}{p_{\mathcal{F}}(x_i, y_i)|\boldsymbol{J}_{X \to Y}|} \right) \ , \quad (5)$$

where $(\boldsymbol{x}, \boldsymbol{n}_Y) = \{(x_i, n_{Y,i})\}_{i=1,\dots,T}$ is an iid sample of size T, $p_{\mathcal{F}}(x, y)$ is the density induced by the SCM with model class $\mathcal{F}$, and $|\boldsymbol{J}_{X \to Y}|$ is the absolute value of the determinant of $\boldsymbol{J}_{X \to Y}$. To arrive at Eq. (5), we exploit that

$$p(x, n_Y) = p_{\mathcal{F}}(x, y)|\boldsymbol{J}_{X \to Y}| \ .$$

Zhang et al. (2015) note that for ANMs, $|\boldsymbol{J}_{X \to Y}| = 1$ can be ignored. For LSNMs, however, $|\boldsymbol{J}_{X \to Y}|$ evaluated at (x, n_Y) is equal to $|g(x)|$ and thus, cannot be ignored.

If $p_{\mathcal{F}}(x, y)$ is induced by a Gaussian LSNM, the conditional log-density of y given x as defined in Eq. 2 already

includes the scaling through $|g(x)|$ since it can be written as $\log p(n_Y|\boldsymbol{\theta}) - \log |g(x)|$. Moreover, $p(x)$ and $p_{\mathcal{F}}(x, y)$ are independent of $\boldsymbol{\theta}$. The mutual information in Eq. (5) is thus minimized as $\log p(y|x, \boldsymbol{\theta})$ is maximized. Moreover, Zhang et al. (2015) show that

$$\mathrm{I}(\boldsymbol{x}, \hat{\boldsymbol{n}}_Y | \hat{\boldsymbol{\theta}}_T) \xrightarrow{p} 0$$

as $T \to \infty$, for any consistent MLE $\hat{\boldsymbol{\theta}}_T$. For LSNMs our estimator derived in Sec. 3 fulfills this condition.

4.1. Likelihood-Based Inference

Cause-effect inference via maximum likelihood or related measures such as through the Minimum Description Length principle or Bayesian estimators has proven to be a successful approach for cause-effect inference (Mooij et al., 2010; Bühlmann et al., 2014; Marx & Vreeken, 2017). The general approach is to compare the log-likelihood estimate (or a related score) for the presumed causal direction $X \to Y$,

$$\ell_{X \to Y}(\hat{\boldsymbol{\theta}}_T) = \sum_{i=1}^{T} \log \left(p(x_i)p\left(y_i|x_i, \hat{\boldsymbol{\theta}}_T\right) \right) \ , \quad (6)$$

to the corresponding score in the $Y \to X$ direction, i.e, $\ell_{Y \to X}(\hat{\boldsymbol{\xi}}_T)$.[4] Thus, we estimate that X causes Y if $\ell_{X \to Y} > \ell_{Y \to X}$, Y causes X if $\ell_{X \to Y} < \ell_{Y \to X}$, and do not decide if both terms as equal.

When closely inspecting Eq. (6), we observe—similar to the ANM case (Zhang et al., 2015)—that comparing $^1/_T \ell_{X \to Y}$ to $^1/_T \ell_{Y \to X}$ for LSNMs is equivalent to comparing $\mathrm{I}(\boldsymbol{x}, \hat{\boldsymbol{n}}_Y | \hat{\boldsymbol{\theta}}_T)$ to $\mathrm{I}(\boldsymbol{y}, \hat{\boldsymbol{n}}_X | \hat{\boldsymbol{\xi}}_T)$ since $p_{\mathcal{F}}(x, y)$ is a constant appearing on both sides. Thus, maximum likelihood can be used to identify cause and effect for an LSNM, whenever the model is identifiable as discussed in Sec. 2 and the estimator is consistent (Sec. 3). We formalize these points in Theorem 3, which is adapted from Zhang et al. (2015).

Theorem 3 *[Zhang et al. (2015)] Let $P_{\mathcal{F}}(X, Y)$ be the joint distribution induced by an LSNM according to Def. 1 with $N_Y \sim \mathcal{N}(0, 1)$, parametrization $\boldsymbol{\theta}$, and consistent MLE $\hat{\boldsymbol{\theta}}_T$. Given an iid sample $(\boldsymbol{x}, \boldsymbol{y}) \sim P_{\mathcal{F}}(X, Y)$ of size T,*

$$\lim_{T \to \infty} \ell_{X \to Y}(\hat{\boldsymbol{\theta}}_T) - \ell_{Y \to X}(\hat{\boldsymbol{\xi}}_T) \geq 0 \ ,$$

with equality, if and only if, a backward model parameterized by $\boldsymbol{\xi}$ exists according to Theorem 4, and $\ell_{Y \to X}(\hat{\boldsymbol{\xi}}_T)$ converges to $\ell_{Y \to X}(\boldsymbol{\xi})$ as $T \to \infty$.

Clearly, if we know $p(x)$ and all assumptions in Theorem 3 are fulfilled, our estimator based on non-linear feature maps is suitable to consistently identify cause and effect according to Theorem 3. In practice, we assume a fully Gaussian model, meaning that we also model $p(x)$ as Gaussian. It would, however, be straightforward to use a different prior.

---

[4]We drop the arguments, whenever clear from context.

4.2. Independence-Based Inference

As a second approach, we apply regression with subsequent independence testing (RESIT) (Peters et al., 2014). Due to the additional degree of freedom to fit the scale of an LSNM, we expect our estimators to be beneficial in this setting.

That is, we first estimate the residuals $\hat{N}_Y = \frac{Y - \hat{f}(X)}{\hat{g}(X)}$ using one of the estimators proposed in Sec. 3, and subsequently test whether $\hat{N}_Y$ is independent of the presumed cause X. We repeat these steps for the $Y \to X$ direction and decide for that direction with the larger p-value; i.e., the smaller the p-value, the more evidence for rejecting independence. If both p-values are insignificant, both directions admit an LSNM, and one could decide to not force a decision.

A possible problem with such a RESIT approach is that the estimated residuals may be inherently dependent on the input (Kpotufe et al., 2014; Mooij et al., 2016), when we do not use a suitable estimator. As discussed by Mooij et al. (2016), a regression estimator is suitable, if i) it is L_2 consistent, or ii) it is weakly consistent, but we perform sample splitting—i.e., we estimate the function parameters on one half of the data and estimate the dependence between residuals and potential cause on the other half of the data. Further, Mooij et al. (2016, Thm. 20) proved that Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005) is consistent given such a suitable estimator.[5] Therefore, we chose HSIC for our empirical evaluation.

L_2-consistency is a stronger statement than what we proved in Theorem 2. For our estimator, we would need to show that the Fisher information is sufficiently peaked at $\boldsymbol{\theta}$ to achieve L_2 consistency, which is difficult for arbitrary nonlinear feature maps. Even though, our estimator is suitable for Gaussian LSNMs if we perform sample splitting, we observe a substantially improved performance when using the full dataset for training and independence testing.

4.3. Location-Scale Causal Inference

To instantiate the estimators above, we propose LOCI (**lo**cation-scale **c**ause-effect **i**nference), which follows a simple two-step procedure: We first standardize the data to avoid any bias that might be induced by the scale of the involved variables (Reisach et al., 2021)[6] Then, we use one of the estimators proposed in Sec. 3 to estimate f and g for the $X \to Y$ direction, and h and k for the $Y \to X$ direction. Last, we determine the causal direction, either by choosing the direction with the higher likelihood, as described in Sec. 4.1 (LOCI$_M$), or the one with the larger p-value provided by HSIC, as proposed in Sec. 4.2 (LOCI$_H$).

We expect the likelihood-based approach to perform best for Gaussian ANMs and LSNMs, since it is consistent in these settings, while we expect LOCI$_H$ to have an advantage when the data is non-Gaussian distributed.

5. Related Work

Cause-effect inference is a well-studied problem, and the first identifiability results have been established for linear additive noise models (Shimizu et al., 2006), where the effect Y can be expressed as a linear function of the cause X and an additive non-Gaussian error term. It has been proven that the backward model does not exist for a broad range of ANMs (Hoyer et al., 2009; Peters et al., 2011; Hu et al., 2018; Peters et al., 2017; Bühlmann et al., 2014), and consistency results for ANMs have been derived (Kpotufe et al., 2014). Further, these results on identifiability have been extended to post non-linear causal models (Zhang & Hyvärinen, 2009) and a link to maximum likelihood estimation has been established (Zhang et al., 2015).

A line of research that is not based on assumptions on the SCM builds upon the principle of independent mechanisms and postulates that the mechanism $(p_{Y|X})$ is independent of the cause (p_X), whereas this does not hold for the anti-causal direction (Janzing & Schölkopf, 2010; Peters et al., 2017). Besides being a postulate, the principle of independent mechanism can also be linked to known identifiability results, e.g., to the likelihood approach we presented in Sec. 4. For Gaussian additive noise models for example, $p_{Y|X}$ models the residual distribution which is independent of p_X for correctly specified models. Note, however, that for location-scale noise models, the definition has to be revisited and we cannot consider $p_{Y|X}$ as an independent mechanism, since the independent noise source is equal to $p_{Y|X} \cdot |\boldsymbol{J}_{X \to Y}|$.[7] Approaches motivated by this postulate either try to approximate the involved distributions directly (Janzing et al., 2012; Sgouritsa et al., 2015; Goudet et al., 2018), or aim to approximate its information-theoretic counterpart: the algorithmic independence of conditionals postulate (Janzing & Schölkopf, 2010), which compares the Kolmogorov complexities (Kolmogorov, 1965) of the involved factorizations. Since Kolmogorov complexity is not computable, it is typically approximated via Minimum Description Length or Minimum Message Length (Kalainathan, 2019; Marx & Vreeken, 2017; Mitrovic et al., 2018; Mooij et al., 2010; Mian et al., 2021). In a broader sense, these approaches also link to (penalized) likelihood methods or Bayesian structure learning (Peters & Bühlmann, 2014; Bühlmann et al., 2014; Marx & Vreeken, 2019b; Lorch et al., 2021), which are of increasing interest, especially for graphs beyond the bivariate setting (Zheng et al., 2018; Vowels et al., 2022).

---

[5]These results have also been extended beyond the two-variable case by Pfister et al. (2018).

[6]Since we assume a Gaussian LSNM, the marginals for X and Y cancel when comparing $\ell_{X \to Y}$ to $\ell_{Y \to X}$ after standardization.

[7]A related concept with identifiability results for discrete data is entropic causal inference (Kocaoglu et al., 2017).
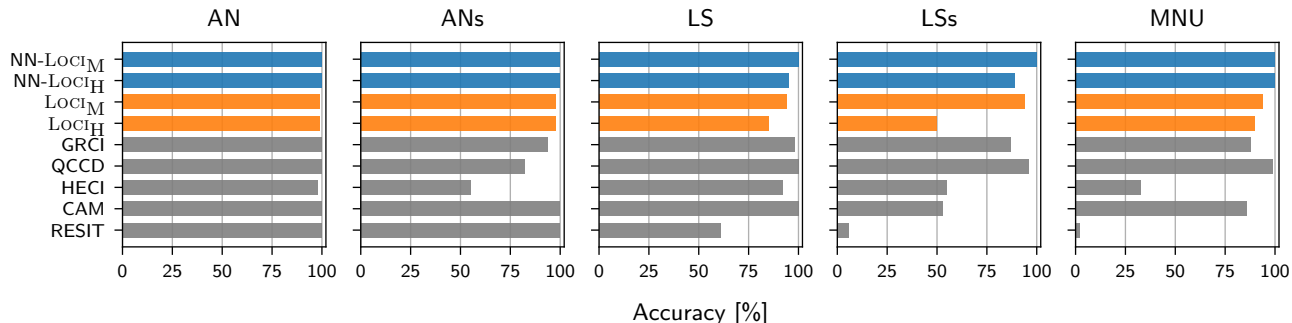
Figure 2. Performance of LOCI with maximum likelihood and HSIC estimator using either neural network (■) or linear regressor (■) in comparison to the baselines (■) on benchmark datasets by Tagasovska et al. (2020). On these datasets, our model is either well-specified or slightly misspecified and therefore LOCI$_M$ performs best. LOCI$_M$ with neural network regressor makes no error on these datasets while other methods designed for similar data misclassify some pairs.

For location-scale noise models, several independent identifiability results exist. HECI (Xu et al., 2022) and FOM (Cai et al., 2020) are two proposals, which extend the identifiability results for ANMs (Blöbaum et al., 2018), which focus on the low-noise regime. In particular, Xu et al. (2022) develop an approach that partitions the domain space and develop a loss based on the geometric mean between error terms, and Cai et al. (2020) suggest a score based on fourth-order moments, which they estimate via heteroscedastic Gaussian process regression. Approaches without identifiability results have also shown strong empirical performance on location-scale benchmark datasets, such as QCCD (Tagasovska et al., 2020), based on non-parametric quantile regression, and estimators for conditional divergences (Fonollosa, 2019; Duong & Nguyen, 2021). Most related to our approach is the concurrent work of (Strobl & Lasko, 2022), who prove Theorem 1 independently of us via a different technique as discussed in Sec. 2. Strobl & Lasko (2022), however, focus on root cause analysis and use a heuristic two stage algorithm, GRCI, to learn the mean and variance via cross validation. Subsequently, they estimate the degree of dependence via mutual information.

In contrast to previous work, we not only provide identifiability results for LSNMs, but also bridge the gap and provide the first consistent estimator for causal LSNMs. Further, we find that our LSNM maximum-likelihood estimator can outperform iterative feasible generalized least-squares (IFGLS, Harvey, 1976), the de-factor standard for heteroscedastic models, especially on small data (cf. Supplementary Material E). Additionally, we provide a neural network-based approach to learn LSNMs, which exhibits state-of-the-art performance on a variety of benchmark tasks.

6. Experiments

We empirically compare LOCI to state-of-the-art bivariate causal inference methods and study the benefit of modelling location-scale noise as well as a post-hoc independence test.

The performance of bivariate causal inference methods is assessed in terms of accuracy and area under the decision rate curve (AUDRC).[8] The accuracy measures the fraction of correctly inferred cause-effect relationships and the AUDRC measures how well the decision certainty indicates accuracy. The certainty is, for example, indicated by the likelihood or p-value difference in both directions. Thus, a high AUDRC indicates that an estimator tends to be correct when it is certain and only incorrect when it is uncertain. Mathematically, given the ground truth direction $t(\cdot)$ and a causal estimator for the direction $f(\cdot)$ on M pairs with ordering $\pi(\cdot)$ according to the estimator's certainty, we have

$$\text{AUDRC} = \frac{1}{M}\sum_{m=1}^{M} \frac{1}{m} \sum_{i=1}^{m} 1_{f(\pi(i))=t(\pi(i))}.$$

That is, we average the accuracy when iteratively adding the pair the estimator is the most certain about. Hence, both accuracy and AUDRC are between 0 and an optimum of 1.

Overall, we find that LOCI performs on par with the best methods for causal pair identification and can achieve perfect accuracy even when the model assumptions are slightly violated. In the case of additive noise models, our estimators perform on par with custom estimators for these cases. In the case of location-scale noise models, our method performs often better than previously proposed estimators for this purpose. We also find that the location-scale noise model estimators we designed greatly improve over corresponding additive noise model estimators with the same structure. Across the 13 benchmarks we consider, our method ranks first and improves over concurrent methods for causal inference of location scale noise models like GRCI (Strobl & Lasko, 2022) and HECI (Xu et al., 2022).

---
[8]We choose AUDRC over AUPR and AUROC, since it weights correctly identified $X \to Y$ pairs in the same way as correctly identified $Y \to X$ pairs and thus avoids an arbitrary selection of true positives and true negatives, which can lead to non-interpretable results on unbalanced datasets (Marx & Vreeken, 2019a, Table 1).

Baselines. We consider CAM (Bühlmann et al., 2014), which is based on homoscedastic maximum likelihood and a corresponding approach based on additional independence tests (Peters et al., 2014, RESIT). Further, we compare to three methods that are also designed for heteroscedasticity: QCCD (Tagasovska et al., 2020), which uses quantile regression, HECI (Xu et al., 2022), which uses binning, and GRCI (Strobl & Lasko, 2022), which uses non-linear regression and cross-validation based scores. In Supplementary Material D, we add CGNN (Goudet et al., 2018) and ICGI (Janzing et al., 2012) as non-LSNM baselines.

Datasets. To assess the performance of our method on datasets where the model assumptions are only slightly violated, we consider the five synthetic datasets proposed by Tagasovska et al. (2020) that consist of additive (AN, ANs), location scale (LS, LSs), and multiplicative (MNU) noise-models. The datasets with suffix 's' use an invertible sigmoidal function making identification more difficult. Further, we assess the performance of our approach on a wide variety of common benchmarks where our assumptions are most likely violated. We consider the Net dataset constructed from random neural networks, the Multi datasets using polynomial mechanisms and various noise settings (including post non-linear noise), and the Cha dataset used for the effect pair challenge (Guyon et al., 2019). Further, we consider the SIM and Tübingen datasets of the benchmark by Mooij et al. (2016).

6.1. Additive, Location-Scale, and Multiplicative Noise

The five datasets proposed by Tagasovska et al. (2020) can be modelled using the location-scale noise model that we assume (Def. 1) and therefore provide an optimal test case for our methods and theoretical results. MNU additionally provides an interesting test-case because it is not generated with Gaussian noise. In Fig. 2, we display the accuracy of our estimators using both a linear model with feature maps and a neural network. We find that our estimators, especially LOCI_M relying on the maximum likelihood, achieve almost-perfect performance on this benchmark. In the following, we compare the performances of the methods in detail.

Comparison to Methods for Additive Noise Models. The direct comparison to CAM (Bühlmann et al., 2014) and RESIT (Peters et al., 2014) is relevant because these two methods rely on additive noise models and can be seen as equivalents to LOCI_M and LOCI_H, respectively. Fig. 2 shows that our method with a neural network estimator, NN-LOCI_M, achieves perfect performance across the five datasets while the method based on feature maps performs only slightly worse. In particular, LOCI_M performs better than the homoscedastic variant CAM and LOCI_H performs significantly better than RESIT. In the case where the as-
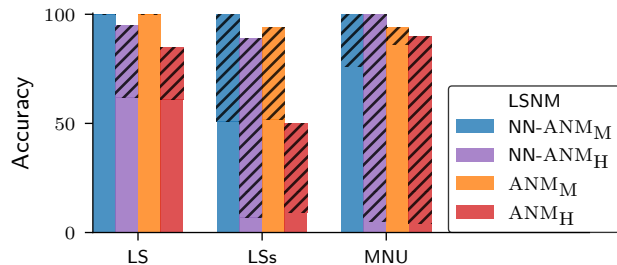

Figure 3. Improvement of using our location-scale noise model (LSNM) estimators (striped area) over corresponding additive noise model (ANM) estimators (solid area) with the same network architectures and feature maps. The ANM estimators use a Gaussian likelihood with fixed observation noise. Only on LS, the likelihood-based ANM estimators are on par with the LSNM variant (ANM_M performing 3% points better than LOCI_M on LS).

sumptions seem to be aligned with our estimators, no independence test is needed and LOCI_M performs better than LOCI_H. In Fig. 3, we further show how the LSNM estimators that we propose improve over an ANM estimator using the same feature maps or neural network architecture. Especially on the three datasets that have complex noise models, LS, LSs, and MNU, our estimators greatly improve the performance. Across all benchmarked datasets, ANM estimators perform the worst on these three and an additional independence test can further decrease the performance below 10% accuracy. Instead, using our LSNM estimators leads to almost perfect accuracy with or without the independence test.

Comparison to location-scale estimators. Although QCCD (Tagasovska et al., 2020), GRCI (Strobl & Lasko, 2022), and HECI (Xu et al., 2022) are designed in particular for heteroscedastic models, they do not achieve perfect performance on this benchmark like NN-LOCI_M. In particular, these methods struggle on additive noise models with sigmoid non-linearity (ANs) and on datasets with the assumed location-scale noise model (LS) while our proposed method, in particular with a flexible neural network, achieves perfect performance across all these cases.

6.2. Performance on other Datatsets

We assess the performance of our method on 8 other benchmark datasets where the LSNM assumptions might be violated. Despite this, we find that the proposed methods perform on par or better than existing state-of-the-art methods for bivariate causal inference. In particular our NN-LOCI_H estimator performs best in aggregate accuracy over all benchmarks followed by GRCI (Strobl & Lasko, 2022). In turn, GRCI (Strobl & Lasko, 2022) performs slightly better in terms of AUDRC followed by our method. Overall, our method performs best on 7 out of the 13 benchmarks
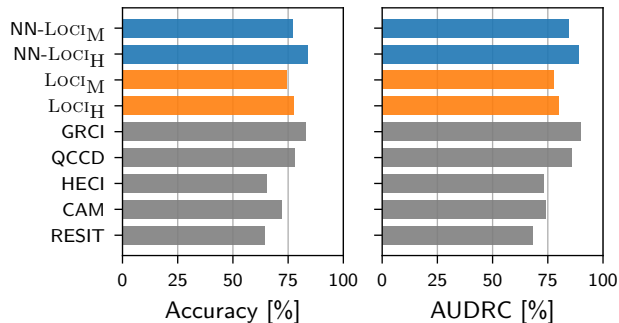
Figure 4. Average accuracy and area under the decision rate curve (AUDRC) over all 13 benchmark datasets. LOCI with a neural network performs best overall in terms of accuracy, GRCI in terms of AUDRC.

considered. The detailed results for all figures can be found in Tables 1 and 2 in the Supplementary Material. In comparison to the five benchmarks where our assumptions hold approximately, the additional independence test greatly helps improve performance on the other 8 benchmark sets.

Overall, the results on various benchmarks suggest that LOCI$_M$ should be used if one expects the Gaussian noise assumptions to hold. If the assumptions are likely violated, LOCI$_H$ should be preferred.

7. Conclusion

We considered the problem of cause-effect inference in location-scale noise models from observational data. We proved that the causal direction is identifiable for LSNMs except for some pathological cases, and proposed two empirical estimators for this setting: a concave feature map-based estimator and a neural network-based approach. For cause-effect inference, we instantiated both variants in a likelihood-based framework, as well as performed subsequent independence testing. The likelihood approach has almost perfect accuracy on benchmark data that approximately meets our assumptions, whereas the independence-based approach is more stable when the data is non-Gaussian. Overall, the neural network instantiation has a slight edge over the feature map-based estimator. The latter, however, outperforms comparable methods such as IFGLS and might be of independent interest. For future work, it would be interesting to inverstigate the non-Gaussian setting in more detail.

Acknowledgements

References

Amemiya, T. *Advanced econometrics*. Harvard university press, 1985.

Blöbaum, P., Janzing, D., Washio, T., Shimizu, S., and Schölkopf, B. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pp. 900–909. PMLR, 2018.

Brown, L. D. *Fundamentals of statistical exponential families: with applications in statistical decision theory*. Ims, 1986.

Bühlmann, P., Peters, J., and Ernest, J. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995.

Cai, R., Ye, J., Qiao, J., Fu, H., and Hao, Z. Fom: Fourth-order moment based causal direction identification on the heteroscedastic data. *Neural Networks*, 124:193–201, 2020.

Cawley, G. C., Talbot, N. L., Foxall, R. J., Dorling, S. R., and Mandic, D. P. Heteroscedastic kernel ridge regression. *Neurocomputing*, 57:105–124, 2004.

Cramér, H. *Mathematical methods of statistics*, volume 43. Princeton university press, 1946.

Duong, B. and Nguyen, T. Bivariate causal discovery via conditional divergence. In *First Conference on Causal Learning and Reasoning*, 2021.

Eilers, P. H. and Marx, B. D. Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2):89–121, 1996.

Fonollosa, J. A. Conditional distribution variability measures for causality detection. In *Cause Effect Pairs in Machine Learning*, pp. 339–347. Springer, 2019.

Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D., and Sebag, M. Learning functional causal models with generative neural networks. *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pp. 39, 2018.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Guyon, I., Statnikov, A., and Batu, B. B. *Cause effect pairs in machine learning*. Springer, 2019.

Harvey, A. C. Estimating regression models with multiplicative heteroscedasticity. *Econometrica: Journal of the Econometric Society*, pp. 461–465, 1976.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 689–696, 2009.

Hu, S., Chen, Z., Partovi Nia, V., CHAN, L., and Geng, Y. Causal inference and mechanism clustering of a mixture of additive noise models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 5212–5222, 2018.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Technology*, 56(10):5168–5194, 2010.

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B., and Schölkopf, B. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

Kalainathan, D. *Generative Neural Networks to infer Causal Mechanisms: algorithms and applications*. PhD thesis, Université Paris-Saclay, 2019.

Khemakhem, I., Monti, R., Leech, R., and Hyvarinen, A. Causal autoregressive flows. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3520–3528. PMLR, 13–15 Apr 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kocaoglu, M., Dimakis, A. G., Vishwanath, S., and Hassibi, B. Entropic causal inference. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Kolmogorov, A. Three approaches to the quantitative definition of information. *Problemy Peredachi Informatsii*, 1(1):3–11, 1965.

Kpotufe, S., Sgouritsa, E., Janzing, D., and Schölkopf, B. Consistency of causal inference under the additive noise model. In *International Conference on Machine Learning*, pp. 478–486. PMLR, 2014.

Kulldorff, G. On the conditions for consistency and asymptotic efficiency of maximum likelihood estimates. *Scandinavian Actuarial Journal*, 1957(3-4):129–144, 1957.

Le, Q. V., Smola, A. J., and Canu, S. Heteroscedastic gaussian process regression. In *Proceedings of the 22nd international conference on Machine learning*, pp. 489–496, 2005.

Liang, K.-Y. The asymptotic efficiency of conditional likelihood methods. *Biometrika*, 71(2):305–313, 1984.

Lin, J. Factorizing multivariate function classes. In *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1997.

Lorch, L., Rothfuss, J., Schölkopf, B., and Krause, A. Dibs: Differentiable bayesian structure learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Martens, J. New insights and perspectives on the natural gradient method. *The Journal of Machine Learning Research*, 21(1):5776–5851, 2020.

Marx, A. and Vreeken, J. Telling cause from effect using mdl-based local and global regression. In *2017 IEEE international conference on data mining (ICDM)*, pp. 307–316. IEEE, 2017.

Marx, A. and Vreeken, J. Telling cause from effect by local and global regression. *Knowledge and Information Systems*, 60(3):1277–1305, 2019a.

Marx, A. and Vreeken, J. Identifiability of cause and effect using regularized regression. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 852–861, 2019b.

Mian, O., Marx, A., and Vreeken, J. Discovering fully oriented causal networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Mitrovic, J., Sejdinovic, D., and Teh, Y. W. Causal inference via kernel deviance measures. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6986–6994, 2018.

Mooij, J. M., Stegle, O., Janzing, D., Zhang, K., and Schölkopf, B. Probabilistic latent variable models for distinguishing between cause and effect. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1687–1695, 2010.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

Pearl, J. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Peters, J. and Bühlmann, P. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Identifiability of causal graphs using functional models. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 589–598. AUAI Press, 2011.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31, 2018.

Reisach, A., Seiler, C., and Weichwald, S. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Rothenberg, T. J. Identification in parametric models. *Econometrica: Journal of the Econometric Society*, pp. 577–591, 1971.

Sgouritsa, E., Janzing, D., Hennig, P., and Schölkopf, B. Inference of cause and effect with unsupervised inverse regression. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 38: 847–855, 2015.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7, 2006.

Strobl, E. V. and Lasko, T. A. Identifying patient-specific root causes with the heteroscedastic noise model. *arXiv preprint arXiv:2205.13085*, 2022.

Tagasovska, N., Chavez-Demoulin, V., and Vatter, T. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *International Conference on Machine Learning*, pp. 9311–9323. PMLR, 2020.

Vowels, M. J., Camgoz, N. C., and Bowden, R. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys (CSUR)*, 2022.

Wald, A. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20 (4):595–601, 1949.

Wooldridge, J. M. *Introductory econometrics: A modern approach*. Cengage learning, 2015.

Xu, S., Mian, O., Marx, A., and Vreeken, J. Inferring cause and effect in the presence of heteroscedastic noise. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pp. 647–655. AUAI Press, 2009.

Zhang, K., Wang, Z., Zhang, J., and Schölkopf, B. On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22, 2015.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 31, 2018.

A. Heteroscedastic Regression Algorithms

Here we describe the estimators we use for maximum likelihood. We assume the data are vectors $\boldsymbol{x} \in \mathbb{R}^T$ and observations $\boldsymbol{y} \in \mathbb{R}^T$. Further, we place an uninformative Gaussian prior on the parameter vector $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \delta^{-1}\boldsymbol{I})$. Assuming conditional independence, the log joint with parameters $\boldsymbol{w}$ is

$$\log p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x}) = \sum_{i=1}^{T} \log p(y_i|x_i, \boldsymbol{w}) + \log \mathcal{N}(\boldsymbol{w}; \boldsymbol{0}, \delta^{-1}\boldsymbol{I}) \ .$$

This ensures that the log joint $\log p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x})$ is strictly concave for our concave estimator and helps with optimization. It can further be used to adjust the complexity of the model by adapting δ. In our case, we use a very small δ with value 10^{-6} which does not regularize but rather helps inverting the Hessian for closed-form updates described below.

A.1. Concave Linear Heteroscedastic Regression

Optimizing the parameters of the concave heteroscedastic regression model with fixed feature maps can partly be done in a closed-form and partly requires gradient-based optimization. The parameter $\boldsymbol{w}_1$ that controls the first natural parameter through $\eta_1(x) = \boldsymbol{\psi}(x)^\mathsf{T}\boldsymbol{w}_1$ can be optimized in closed-form. We define $\boldsymbol{\Psi} = [\boldsymbol{\psi}(x_1), \ldots, \boldsymbol{\psi}(x_T)]^\mathsf{T} \in \mathbb{R}^{T \times D}$ as the design matrix. Differentiating the log-likelihood with respect to $\boldsymbol{w}_1$, we have

$$\nabla_{\boldsymbol{w}_1} \log p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x}) = \boldsymbol{\Psi}^\mathsf{T}\left(\boldsymbol{y} - \boldsymbol{\alpha} \circ \boldsymbol{\Psi}\boldsymbol{w}_1\right) - \delta\boldsymbol{w}_1,$$

where $\boldsymbol{\alpha} = -\left[\frac{1}{2\eta_2(x_1)}, \ldots, \frac{1}{2\eta_2(x_T)}\right]^\mathsf{T}$ denotes variances for each data point. Therefore, this is easily recognized as a weighted least-squares problem that can be solved as

$$\boldsymbol{w}_1 \leftarrow (\boldsymbol{\Psi}^\mathsf{T}\mathrm{diag}(\boldsymbol{\alpha})\boldsymbol{\Psi} + \delta\boldsymbol{I})^{-1}\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{y},$$

where $\mathrm{diag}(\boldsymbol{\alpha})$ is a $T \times T$ diagonal matrix with entries $\boldsymbol{\alpha}$. While the objective is also concave in $\boldsymbol{w}_2$ that controls the second natural parameter $\eta_2(x)$, there is no closed-form solution available. Therefore, we optimize this part of the objective using the L-BFGS optimizer (Byrd et al., 1995) with the additional constraint that $\boldsymbol{w}_2 \geq 0$. We then fit the model alternatingly, similar to FGLS, by first using the closed-form solution for $\boldsymbol{w}_1$ and then optimizing the second parameter $\boldsymbol{w}_2$ for multiple steps using L-BFGS. We iterate until the objective value does not improve anymore.

A.2. Neural Network Heteroscedastic Regression

In general, it is expected that our model is misspecified or that we do not know the underlying feature maps. To deal with this case, we use neural networks that can automatically learn the feature maps specifically per problem instead of relying on splines or more traditional feature maps. In contrast to common implementations of neural networks for heteroscedastic regression, which parameterize the mean and variance, we employ the natural parametrization in line with our model based on feature maps. With neural network $\boldsymbol{f} : \mathbb{R} \times \mathbb{R}^D \to \mathbb{R}^2$ we have

$$\eta_1(x) = f_1(x, \boldsymbol{w}) \ \text{ and } \ \eta_2(x) = -\tfrac{1}{2}\exp(f_2(x, \boldsymbol{w})),$$

which uses an additional exponential link function to ensure positivity while maintaining differentiability. Further, above model is sensible because for a single layer a zero-mean Gaussian prior would cause a prior predictive mode at unit variance and hence match the data after standardization. To optimize the model, we use the Adam optimizer (Kingma & Ba, 2014).

B. Theoretical Results

Theorem 1 *Assume the data is such that a location-scale noise model can be fit in both directions, i.e.,*

$$Y = f(X) + g(X)N_Y, \quad X \perp\!\!\!\perp N_Y$$
$$X = h(Y) + k(Y)N_X, \quad Y \perp\!\!\!\perp N_X.$$

Let $\nu_1(\cdot)$ and $\nu_2(\cdot)$ be the twice differentiable log densities of Y and N_X respectively. For compact notation, define

$$\nu_{X|Y}(x|y) = \log\left(p_{X|Y}(x|y)\right)$$

$$= \log\left(p_{N_X}\left(\frac{x - h(y)}{k(y)}\right)/k(y)\right)$$

$$= \nu_2\left(\frac{x - h(y)}{k(y)}\right) - \log(k(y)) \quad \text{and}$$

$$G(x,y) = g(x)f'(x) + g'(x)[y - f(x)].$$

Assume that $f(\cdot)$, $g(\cdot)$, $h(\cdot)$, and $k(\cdot)$ are twice differentiable. Then, the data generating mechanism must fulfill the following PDE for all x, y with $G(x,y) \neq 0$.

$$0 = \nu_1''(y) + \frac{g'(x)}{G(x,y)}\nu_1'(y) + \frac{\partial^2}{\partial y^2}\nu_{X|Y}(x|y) + \tag{1}$$

$$\frac{g(x)}{G(x,y)}\frac{\partial^2}{\partial y \partial x}\nu_{X|Y}(x|y) + \frac{g'(x)}{G(x,y)}\frac{\partial}{\partial y}\nu_{X|Y}(x|y).$$

Proof: We follow the proof technique of (Zhang & Hyvärinen, 2009), i.e., we build upon the linear separability of the logarithm of the joint density of independent random variables. That is, for a set of independent random variables whose joint density is twice differentiable, the Hessian of the logarithm of their density function is diagonal everywhere (Lin, 1997). We first define the joint distribution $p(x, n_Y)$ via the change of variable formula, then derive the Hessian of its logarithm, and lastly, derive an PDE which is necessary to hold such that an inverse model can exist.

We define the change of variables from $\{x, n_Y\}$ to $\{y, n_X\}$

$$y = f(x) + g(x)n_Y,$$
$$n_X = [x - h(y)]/k(y).$$

The according Jacobian matrix amounts to

$$\begin{pmatrix} \frac{\partial y/\partial x}{k(y)} - \partial y/\partial x \frac{k(y)h'(y) + [x - h(y)]k'(y)}{k(y)^2} & -g(x)\frac{k(y)h'(y) + [x - h(y)]k'(y)}{k(y)^2} \end{pmatrix},$$

with absolute determinant $g(x)/k(y)$ such that

$$p(x, n_Y) = \frac{g(x)}{k(y)}p(y, n_X).$$

Under independence it holds

$$\frac{\partial^2}{\partial x \partial n_Y}\log(p(x, n_Y)) = 0 \quad \text{such that}$$

$$\frac{\partial^2}{\partial x \partial n_Y}\log\left(\frac{g(x)}{k(y)}p(y, n_X)\right) = \frac{\partial^2}{\partial x \partial n_Y}[\nu_1(y) + \nu_2(n_X) + \log(g(x)) - \log(k(y))] = 0.$$

Evaluating this quantity and dividing by $(\partial y/\partial x)(\partial y/\partial n_Y)$ leads to

$$\nu_1''(y) + \nu_1'(y)\left[\frac{g'(x)}{G(x, n_Y)}\right] + \nu_2''(n_X)\left[\frac{h'(y) + k'(y)n_X}{k(y)^2}\left[h'(y) + k'(y)n_X - \frac{g(x)}{G(x, n_Y)}\right]\right] +$$

$$\nu_2'(n_X)\left[\frac{2\left[n_X k'(y)^2 + h'(y)k'(y)\right]}{k(y)^2} - \frac{h''(y) + n_X k''(y)}{k(y)} - \frac{h'(y)g'(x) + n_X k'(y)g'(x)}{k(y)G(x, n_Y)} - \right.$$

$$\left.\frac{g(x)k'(y)}{k(y)^2 G(x, n_Y)}\right] + \frac{k'(y)^2 - k(y)k''(y)}{k(y)^2} - \frac{g'(x)k'(y)}{k(y)G(x, n_Y)} = 0, \tag{7}$$

where

$$G\left(x, n_Y\right) = \frac{\partial y}{\partial x} \frac{\partial y}{\partial n_Y} = g\left(x\right)\left[f'\left(x\right) + g'\left(x\right) n_Y\right].$$

Assuming injectivity of $f\left(\cdot\right)$ and positivity of $g\left(\cdot\right)$, $G\left(x, n_Y\right)$ can only be 0 on a set of measure 0. Finally, plugging in all the definitions in Eq. (1) and taking the derivatives, one finds that Eq. (1) is identical to Eq. (7). $\qquad\square$

Lemma 1 *With natural parameters modelled as $\eta_1(x) = \psi(x)^{\mathsf{T}} w_1$ and $\eta_2(x) = -\phi(x)^{\mathsf{T}} w_2$, the Gaussian log-likelihood function in Eq. (2) is jointly concave in w_1, w_2.*

Proof: To prove concavity of the model, we show that the Hessian of the loss w.r.t. parameters w_1, w_2 is negative semidefinite. We write $\boldsymbol{\eta}(x) = [\eta_1(x), \eta_2(x)]^{\mathsf{T}} \in \mathbb{R}^2$ for the natural parameter vector and thus have the log-likelihood

$$\log p(y|x, \boldsymbol{\eta}) = -\frac{1}{2}\log\left(2\pi\right) + \boldsymbol{\eta}(x)^{\mathsf{T}} \begin{bmatrix} y \\ y^2 \end{bmatrix} + \frac{\eta_1(x)^2}{4\eta_2(x)} + \frac{1}{2}\log(-2\eta_2(x)) .$$

Further, we concatenate both parameters and feature maps so that we have $w = [w_1^{\mathsf{T}} w_2^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{R}^{2D}$ and $\boldsymbol{\Phi}(x) \in \mathbb{R}^{2D \times 2}$ is a block-diagonal concatenation of feature maps such that the first column is $\psi(x)$ followed by zeros and the second column is D zeros followed by $-\phi(x)$, i.e.,

$$\boldsymbol{\Phi}(x) = \begin{pmatrix} \psi(x) & \mathbf{0} \\ \mathbf{0} & -\phi(x) \end{pmatrix} .$$

We then have $\boldsymbol{\eta}(x) = \boldsymbol{\Phi}(x)^{\mathsf{T}} w$ for our natural parameters. By the chain rule, the Hessian w.r.t. parameters w can be written as

$$\nabla_{\boldsymbol{w}}^2 \log p(y|x, \boldsymbol{w}) = [\nabla_{\boldsymbol{w}} \boldsymbol{\eta}(x)]^{\mathsf{T}} [\nabla_{\boldsymbol{\eta}}^2 \log p(y|x, \boldsymbol{\eta})][\nabla_{\boldsymbol{w}} \boldsymbol{\eta}(x)] + [\nabla_{\boldsymbol{w}}^2 \boldsymbol{\eta}(x)]^{\mathsf{T}} [\nabla_{\boldsymbol{\eta}} \log p(y|x, \boldsymbol{\eta}(x))]$$
$$= \boldsymbol{\Phi}(x)[\nabla_{\boldsymbol{\eta}}^2 \log p(y|x, \boldsymbol{\eta})]\boldsymbol{\Phi}(x)^{\mathsf{T}},$$

because the second derivative of $\boldsymbol{\eta}(x) = \boldsymbol{\Phi}(x)^{\mathsf{T}} w$ w.r.t. w is zero due to linearity thus eliminating the second summand. Because negative definiteness of $\nabla_{\boldsymbol{\eta}}^2 \log p(y|x, \boldsymbol{\eta})$ implies negative semidefiniteness for any matrix $\boldsymbol{\Phi}(x)\nabla_{\boldsymbol{\eta}}^2 \log p(y|x, \boldsymbol{\eta})\boldsymbol{\Phi}(x)^{\mathsf{T}}$, it only remains to show that $\nabla_{\boldsymbol{w}}^2 \log p(y|x, \boldsymbol{w})$ is indeed negative definite. While negative definiteness already follows from the natural parameterisation, the Hessian is given by:

$$\nabla_{\boldsymbol{\eta}}^2 \log p(y|x, \boldsymbol{\eta}) = \begin{pmatrix} \frac{1}{2\eta_2(x)} & -\frac{\eta_1(x)}{2\eta_2(x)^2} \\ -\frac{\eta_1(x)}{2\eta_2(x)^2} & \frac{\eta_1(x)^2}{2\eta_2(x)^3} - \frac{1}{2\eta_2(x)^2} \end{pmatrix} .$$

To show that this matrix is negative definite, we use the determinant criterion. This requires that odd upper determinants be negative and even ones positive. The upper left determinant is negative since $\frac{1}{2\eta_2(x)}$ is negative due to $\eta_2(x) < 0$. The determinant of the entire matrix is given by $-\frac{1}{4\eta_2(x)^3}$ and is positive due to $\eta_2(x) < 0$. This shows that the log-likelihood Hessian w.r.t. η is negative definite and concludes the proof that the log-likelihood is concave in w. $\qquad\square$

Theorem 2 *Let $\{(y_i, x_i)\}_{i=i,\ldots,T}$ be an iid sample with conditional density $p(y|x_i; w^*)$ as defined through Eqs. (2) and (3), with $w^* \in \mathbb{R}^D \times \mathbb{R}_+^D$, and correctly specified (non-linear) feature maps $\psi(x) \in \mathbb{R}^D$ and $\phi(x) \in \mathbb{R}_+^D$. Further, there exists a $T_0 > 0$ for which Assumption 1 holds, and suppose the usual smoothness criteria hold:*

  i) *Derivatives up to second order can be passed under the integral sign in $\int dP(y|x, w)$.*

  ii) *Third derivatives $\nabla^3 \log p(y|x, w)$ are bounded by a function $M(y|x)$ on $\mathbb{R}^D \times \mathbb{R}_+^D$, s.t.*

$$\sup_{\theta \in \mathbb{R}^D \times \mathbb{R}_+^D} |\nabla^3 \log p(y|x, w)_{jkl}| \leq M(y|x)$$

  for all j, k, l, and $\mathbb{E}[M(y|x)] < \infty$.

Then, for $T \geq T_0$ the maximum likelihood estimate $\hat{w}_T$ is weakly consistent, i.e., for $T \to \infty$, $\hat{w}_T \overset{p}{\to} w^*$. Further, $\hat{w}_T$ is asymptotically normal in a sense that $\sqrt{T}(\hat{w}_T - w^*) \overset{d}{\to} \mathcal{N}(0, I(w^*)^{-1})$.

14

Proof: We first establish that the true parameter $\boldsymbol{w}^*$ is identifiable. Since the density $p(y|x, \boldsymbol{w}^*)$ is part of the exponential family, the parameter space $\mathbb{R}^D \times \mathbb{R}^D_+$ is a convex set, and for $T \geq T_0$, $I(\boldsymbol{w}) > 0$ for all $\boldsymbol{w} \in \mathbb{R}^D \times \mathbb{R}^D_+$ (by Assumption 1), it follows that any $\boldsymbol{w} \in \mathbb{R}^D \times \mathbb{R}^D_+$ is globally identifiable as proven by Rothenberg (1971)[Theorem 3].

Since there exists a $T \geq T_0$, so that $I(\boldsymbol{w}) > 0$ for any $\boldsymbol{w} \in \mathbb{R}^D \times \mathbb{R}^D_+$, the main condition for Cramér's consistency proof is also fulfilled (Cramér, 1946). Combined with the smoothness conditions i) and ii), we fulfill all criteria according to Cramér (1946) resp. Liang (1984), who laid out the assumptions for conditional log-likelihoods such that the claim holds. □

Theorem 3 *[Zhang et al. (2015)] Let $P_\mathcal{F}(X, Y)$ be the joint distribution induced by an LSNM according to Def. 1 with $N_Y \sim \mathcal{N}(0, 1)$, parametrization $\boldsymbol{\theta}$, and consistent MLE $\hat{\boldsymbol{\theta}}_T$. Given an iid sample $(\boldsymbol{x}, \boldsymbol{y}) \sim P_\mathcal{F}(X, Y)$ of size T,*

$$\lim_{T \to \infty} \ell_{X \to Y}(\hat{\boldsymbol{\theta}}_T) - \ell_{Y \to X}(\hat{\boldsymbol{\xi}}_T) \geq 0 \, ,$$

with equality, if and only if, a backward model parameterized by $\boldsymbol{\xi}$ exists according to Theorem 4, and $\ell_{Y \to X}(\hat{\boldsymbol{\xi}}_T)$ converges to $\ell_{Y \to X}(\boldsymbol{\xi})$ as $T \to \infty$.

Proof: From Eq. (2) and Eq. (5), it follows that for LSNMs,

$$\ell_{X \to Y}(\hat{\boldsymbol{\theta}}_T) = \sum_{i=1}^{T} p_\mathcal{F}(x_i, y_i) - T \cdot \mathrm{I}(\boldsymbol{x}, \hat{\boldsymbol{n}}_Y | \hat{\boldsymbol{\theta}}_T)$$

$$\ell_{Y \to X}(\hat{\boldsymbol{\xi}}_T) = \sum_{i=1}^{T} p_\mathcal{F}(x_i, y_i) - T \cdot \mathrm{I}(\boldsymbol{y}, \hat{\boldsymbol{n}}_X | \hat{\boldsymbol{\xi}}_T) \, .$$

Thus, it suffices to show that

$$\lim_{T \to \infty} \mathrm{I}(\boldsymbol{x}, \hat{\boldsymbol{n}}_Y | \hat{\boldsymbol{\theta}}_T) - \mathrm{I}(\boldsymbol{y}, \hat{\boldsymbol{n}}_X | \hat{\boldsymbol{\xi}}_T) \leq 0 \, ,$$

with equality, if and only if, a backward model parameterized by $\boldsymbol{\xi}$ exists according to Theorem 4, and $\ell_{Y \to X}(\hat{\boldsymbol{\xi}}_T)$ converges to $\ell_{Y \to X}(\boldsymbol{\xi})$ as $T \to \infty$. By consistency of $\hat{\boldsymbol{\theta}}_T$, $\lim_{T \to \infty} I(\boldsymbol{x}, \hat{\boldsymbol{n}}_Y | \hat{\boldsymbol{\theta}}_T) = 0$ (Zhang et al., 2015). Further, if and only if a backward model exists s.t. $Y \perp\!\!\!\perp N_X$, the mutual information in the backward direction is zero. As a consequence, the empirical estimator approaches zero, if and only if a backward model exist and $\ell_{Y \to X}(\hat{\boldsymbol{\xi}}_T)$ approaches $\ell_{Y \to X}(\boldsymbol{\xi})$ as $T \to \infty$. Otherwise $\lim_{T \to \infty} \mathrm{I}(\boldsymbol{y}, \hat{\boldsymbol{n}}_X | \hat{\boldsymbol{\xi}}_T) > 0$. □

In the following, we state the theoretical result by Khemakhem et al. (2021) on Gaussian LSNMs. Note that we slightly changed the theorem as the original version has a typo in the definition of g and k.

Theorem 4 (Khemakhem et al. (2021)) *Assume the data follows the model in Def. 1 with N_Y standard Gaussian, $N_Y \sim \mathcal{N}(0, 1)$. If a backward model exists, i.e.*

$$X = h(Y) + k(Y)N_X$$

where $N_X \sim \mathcal{N}(0, 1)$, $N_X \perp\!\!\!\perp Y$ and $k > 0$, then one of the following scenarios must hold:

1. *$(g, f) = \left(\frac{1}{\sqrt{Q}}, \frac{P}{Q}\right)$ and $(k, h) = \left(\frac{1}{\sqrt{Q'}}, \frac{P'}{Q'}\right)$ where Q, Q' are polynomials of degree two, $Q, Q' > 0$, P, P' are polynomials of degree two or less, and p_X, p_Y are strictly log-mix-rational-log. In particular, $\lim_{-\infty} g = \lim_{+\infty} g = 0^+$, $\lim_{-\infty} f = \lim_{+\infty} f < \infty$, similarly so for k, h, and f, g, h, k are not invertible.*

2. *g, k are constant, f, g are linear and p_X, p_Y are Gaussian densities.*

C. Societal Impact

Causal inference in general has the potential to greatly improve the reliability of learning systems by uncovering causal relationships instead of, potentially spurious, correlations. Our work contributes to this endeavour and we do not anticipate any negative societal impacts because our work is mostly of theoretical nature.

D. Experimental Protocol and Results

All considered datasets were standardized to zero mean and unit variance for our methods, and use the recommended pre-processing for the baselines. We use the scores of the estimators provided and only if they are strictly greater (or lower for some methods) in the right direction, a given pair is counted as correct. For the baselines we use the standard settings of the respective papers introducing them. Except for standardization, the datasets are not adjusted except for the Tübingen dataset, where we remove the 6 multivariate and 3 discrete pairs. For the overall performance in Fig. 4, we weight each dataset the same irrespective of the number of pairs.

For our estimators LOCI$_M$ and LOCI$_H$, we use spline feature maps (Eilers & Marx, 1996) of order 5 and with 25 knots implemented in `scikit-learn` (Pedregosa et al., 2011). We run our maximum likelihood estimator LOCI$_M$ for up to 100 steps or until the likelihood change in a step is below 10^{-6}. For the independence-test-based estimator, we compute the residuals r of the estimator and standardize them using the estimate of the standard deviation s with $r = \frac{y-\hat{\mu}}{\hat{\sigma}}$ and then test for independence of r and the corresponding input x. We predict based on the p-value of the test in both directions.

For the neural network based estimators, NN-LOCI$_M$ and NN-LOCI$_H$, we use a neural network with a single hidden layer of width 100 and `TanH` activation function. The first output of the network is unrestricted and the second output applies an exponential function to ensure a positive output as described in Eq. 3 and App. A.2. The parameters are optimized by full-batch gradient-descent on the log-likelihood using `Adam` (Kingma & Ba, 2014) with initial learning rate 10^{-2} that is decayed to 10^{-6} using a cosine learning rate schedule for $5\,000$ steps.

All numbers displayed in the main text figures are organized below in Tables 1 and 2 with additional baselines IGCI (Janzing et al., 2012) and CGNN (Goudet et al., 2018).

	AN	ANs	LS	LSs	MNU	SIM	SIMc	SIMln	SIMG	Tue	Cha	Net	Multi
NN-LOCI$_M$	**100**	**100**	**100**	**100**	**100**	48	50	79	78	57	43	76	72
NN-LOCI$_H$	**100**	**100**	95	89	**100**	**79**	**83**	72	78	60	**72**	**87**	78
LOCI$_M$	99	98	94	94	93	52	48	77	74	52	46	75	66
LOCI$_H$	99	98	85	53	90	75	76	73	**81**	56	70	84	66
GRCI	**100**	94	98	87	88	77	77	77	70	**82**	70	85	77
QCCD	**100**	82	**100**	96	99	62	72	80	64	77	54	80	51
HECI	98	55	92	55	33	49	55	65	56	71	57	72	91
CGNN	99	90	98	90	97	48	57	71	80	66	61	69	71
IGCI	20	35	46	34	11	37	45	51	53	68	55	55	**92**
CAM	**100**	**100**	**100**	53	86	57	60	**87**	**81**	58	47	78	35
RESIT	**100**	**100**	61	6	2	77	82	**87**	78	57	**72**	78	37

Table 1. Accuracies of all methods on all benchmark data sets considered. These data were used for the barplots in the main text.

	AN	ANs	LS	LSs	MNU	SIM	SIMc	SIMln	SIMG	Tue	Cha	Net	Multi
NN-LOCI$_M$	**100**	**100**	**100**	**100**	**100**	60	63	**95**	89	66	47	86	93
NN-LOCI$_H$	**100**	**100**	99	97	**100**	89	**93**	86	**93**	56	71	**97**	77
LOCI$_M$	98	95	88	86	90	68	56	90	87	45	55	84	75
LOCI$_H$	**100**	96	92	54	92	84	85	88	91	47	68	93	51
GRCI	**100**	**100**	**100**	95	97	**90**	92	92	88	73	71	96	74
QCCD	**100**	91	**100**	**100**	**100**	71	83	92	76	**84**	61	94	63
HECI	**100**	63	99	72	20	59	64	86	73	78	57	84	**99**
CGNN	**100**	99	**100**	98	**100**	59	66	89	**93**	77	63	81	91
IGCI	18	35	60	49	1	34	41	51	63	74	58	61	**99**
CAM	**100**	**100**	**100**	36	76	68	69	87	88	70	41	85	40
RESIT	**100**	**100**	69	4	0	75	84	83	71	71	**83**	81	68

Table 2. Area under the decision rate curve (AUDRC) of all the estimators used on all benchmark data sets. This data was used for the overall performance displayed in Fig. 4.

E. Performance Comparison of LSNM Estimators

Our LSNM estimator that can achieve consistency under the right assumptions is different to the standard method commonly used for such estimation. Because we use the natural exponential family parametrization of the Gaussian log-likelihood, we have concavity in the linear model parameters while different common formulations do not have this property. The de-facto standard is iterative feasible generalized least-squares (IFGLS) estimator that regresses to the mean and the squared residuals in an alternating fashion (Harvey, 1976; Cawley et al., 2004). Such estimators are mostly developed and used in the context of econometrics (Amemiya, 1985; Wooldridge, 2015) and, to the best of our knowledge, an approach based on natural parameters of the Gaussian location-scale noise model has only been proposed in the context of Gaussian process regression (Le et al., 2005). In the following, we empirically compare our method to IFGLS.

In Fig. 5 and 6, we compare both estimators on a simple sinusoidal example with $x \sim \mathcal{U}[-4\pi, 4\pi]$ and $y \sim \mathcal{N}(\sin(x), 0.1(4\pi - |x|) + 0.2)$ so that the observation noise is small at the borders and large in the middle of the problem domain. We increase the number of samples T from 100 to 10000 and report the KL-divergence from the estimated predictive density q_{est} to the true density p averaged over a grid of 10^4 points from the left to the right boundary of the dataset (-4π to 4π). We repeat this procedure for 100 times and, due to numerical outliers, show the median performance of both estimators. In Fig. 5, we use the common heuristic of setting the number of knots of the spline features to $\sqrt{T}$ and in 6, we set it to $\frac{T}{10}$ leading to overly complex features that make fitting harder.

Overall, we observe that the proposed estimator is more robust for smaller datasets for both strategies of selecting the number of features. Further, IFGLS fails catastrophically for too many complex features due to overfitting of the mean and thus iteratively also the variance. Our estimator does not seem to suffer from this problem. We hypothesize that this is due to the jointly concave objective proposed. However, it is apparent from Fig. 5 that both estimators behave asymptotically the same. This shows overall that our proposed estimator performs appropriately and, as can be seen on the right hand side of both figures, can successfully fit mean and variance simultaneously. We hypothesize that both estimators work equally well given optimal feature maps and leave a thorough investigation of the proposed estimator for future work.
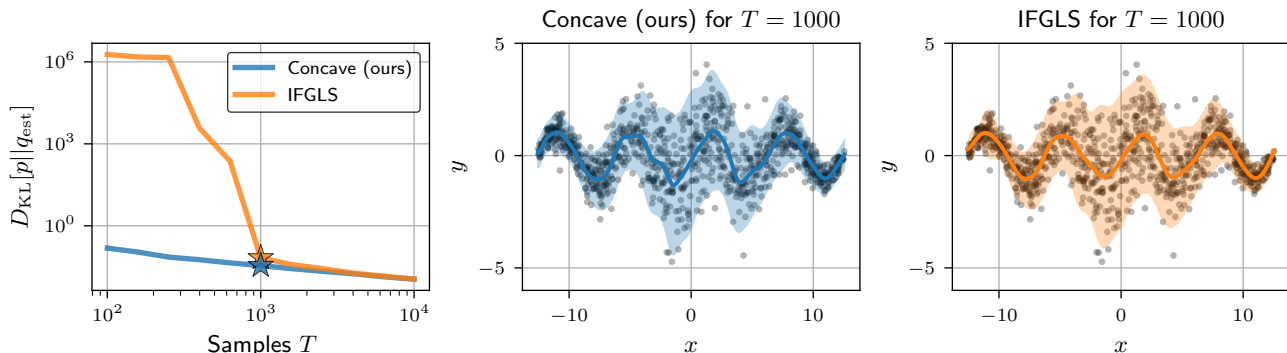

Figure 5. Performance of our maximum likelihood estimator using spline features of order 5 and the common heuristic of knots $= \sqrt{T}$. In comparison to iterative feasible generalized least-squares (IFGLS in ▬), our concave estimator (▬) performs significantly better with fewer samples although both behave asymptotically the same. The right two figures show the fits at $T = 1000$.
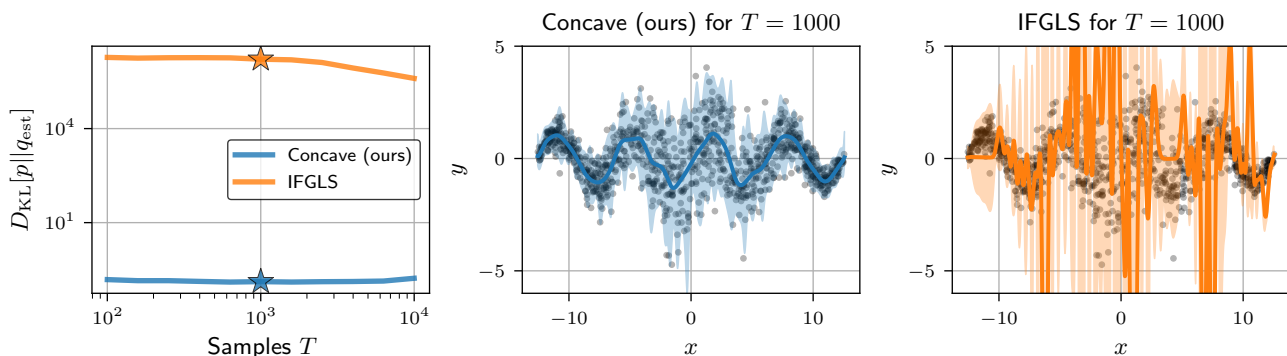

Figure 6. Performance of our maximum likelihood estimator using spline features of order 5 and more features than common with knots $= \frac{T}{10}$ with T samples. Additional complexity makes IFGLS greatly overfit and prevents it from converging while the concave estimator still works somewhat reliably. On the right, the two fits at $T = 1000$ of both methods are shown.