Appendix

A Identifiability of the affine causal model

Recall the form of the SEM that is defined by an autoregressive affine flow:

$$x_j = e^{s_j(x_{<\pi(j)})} z_j + t_j(x_{<\pi(j)}), \quad j = 1, 2 \tag{9}$$

where π is a permutation that describes the causal ordering.

The proof for additive flows ($s_1 = s_2 = 0$ in equation (9)) and general noise can be found in Hoyer et al. (2009). Theorem 2 below summarizes the two scenarios under which the causal model defined by an affine flow is not identifiable. In particular, if the function t_j in equation (9) linking cause to effect is invertible and non-linear, then none of these scenarios can hold. In addition, the proof of Theorem 2 only requires one of the noise variables to be Gaussian.

Definition 1. *Let $(\alpha, \gamma, \delta, \beta, \alpha_0, \beta_0, \gamma_0, \delta_0) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^2_{>0} \times \mathbb{R}^5$ be a tuple such that one of the following conditions holds:*

- $\alpha > 0$, $\alpha_0^2 < \alpha\delta$ *and* $\beta^2 < 4\alpha\gamma$.
- $\alpha = \beta = \alpha_0 = 0$ *and* $\beta_0^2 < \delta$.

We say that a density p_x of a continuous variable x is log-mix-rational-log *if it has the form:*

$$\log p_x(x) = -\frac{1}{2}\delta x^2 + \delta_0 x + \frac{1}{2}\frac{\left(\alpha_0 x^2 + \beta_0 x + \gamma_0\right)^2}{\alpha x^2 + \beta x + \gamma} - \frac{1}{2}\log(\alpha x^2 + \beta x + \gamma) + const \tag{10}$$

We say that p_x is strictly *log-mix-rational-log if $\alpha > 0$.*

Note that the Gaussian distribution is part of the log-mix-rational-log family, for $\alpha = \beta = \alpha_0 = 0$. If $\alpha \neq 0$, then the log-mix-rational-log family is not part of the exponential family.

Theorem 2. *Assume the data follows the model*

$$y = f(x) + v(x)n \tag{11}$$

where n is a standardized Gaussian independent of x, f and v are twice-differentiable scalar functions defined on $\mathbb{R}$ and $v > 0$.
If a backward model exists, i.e. *the data also follows the same model in the other direction*

$$x = g(y) + w(y)m \tag{12}$$

where m is a standardized Gaussian independent of y and $w > 0$, then one of the following scenarios must hold:

1. $(v, f) = \left(\frac{1}{Q}, \frac{P}{Q}\right)$ *and* $(w, g) = \left(\frac{1}{Q'}, \frac{P'}{Q'}\right)$ *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,* $\lim_{-\infty} v = \lim_{+\infty} v = 0^+$, $\lim_{-\infty} f = \lim_{+\infty} f < \infty$, $\lim_{-\infty} w = \lim_{+\infty} w = 0^+$, $\lim_{-\infty} g = \lim_{+\infty} g < \infty$ *and f, v, g, w are not invertible.*
2. *v, w are constant, f, g are linear and p_x, p_y are Gaussian densities.*

Proof. The log-likelihood of (11), denoted by p_1, is given by

$$\log p_1(x, y) = \log p_x(x) - \frac{1}{2}\left(\frac{y - f(x)}{v(x)}\right)^2 - \log v(x) - \frac{1}{2}\log 2\pi \tag{13}$$

and log-likelihood of (12), denoted by p_2, is given by

$$\log p_2(x, y) = \log p_y(y) - \frac{1}{2}\left(\frac{x - g(y)}{w(y)}\right)^2 - \log w(y) - \frac{1}{2}\log 2\pi \tag{14}$$

If the data follows both models, these are equal:

$$\log p_x(x) - \frac{1}{2}\left(\frac{y - f(x)}{v(x)}\right)^2 - \log v(x) = \log p_y(y) - \frac{1}{2}\left(\frac{x - g(y)}{w(y)}\right)^2 - \log w(y) \tag{15}$$

Denote $\frac{1}{v(x)}$ by $\overline{v}(x)$ and likewise for w. Now, take the derivative of both sides with respect to x:

$$(\log p_x)'(x) - \overline{v}(x)(y - f(x))(y\overline{v}'(x) - (f\overline{v})'(x)) - (\log v)'(x) = -(x - g(y))\overline{w}^2(y) \tag{16}$$

Take the derivative of both sides of this with respect to y:

$$-\overline{v}(x)[2y\overline{v}'(x) - (f\overline{v})'(x) - f(x)\overline{v}'(x)] = -x(\overline{w}^2)'(y) + g'(y)\overline{w}^2(y) + g(y)(\overline{w}^2)'(y) \tag{17}$$

Again, take the derivative of both sides with respect to x:

$$-y(\overline{v}^2)''(x) + [\overline{v}((f\overline{v})' + f\overline{v}')]'(x) = -(\overline{w}^2)'(y) \tag{18}$$

and once more, take the derivative of both sides of this with respect to y:

$$-(\overline{v}^2)''(x) = -(\overline{w}^2)''(y) \tag{19}$$

which is possible only if both sides are constant, which is equivalent to $\overline{v}^2$ and $\overline{w}^2$ being second-order polynomials. In other words,

$$\overline{v}^2(x) = \alpha x^2 + \beta x + \gamma, \quad v^2(x) = \frac{1}{\alpha x^2 + \beta x + \gamma} \tag{20}$$

where the parameters must be such that the $\overline{v}$ is always positive. The same holds for w:

$$\overline{w}^2(y) = \alpha' y^2 + \beta' y + \gamma', \quad w^2(y) = \frac{1}{\alpha' y^2 + \beta' y + \gamma'} \tag{21}$$

Furthermore, equation (18) together with the fact that $(\overline{v}^2)''(x) = \text{const}$ implies that:

$$[\overline{v}((f\overline{v})' + f\overline{v}')]'(x) = [f'\overline{v}^2 + 2f(\overline{v}^2)')]'(x) = (f\overline{v}^2)''(x) = \text{const} \tag{22}$$

or

$$f(x)\overline{v}^2(x) = \alpha_0 x^2 + \beta_0 x + \gamma_0 \tag{23}$$

which means that f has the following form:

$$f(x) = \frac{\alpha_0 x^2 + \beta_0 x + \gamma_0}{\alpha x^2 + \beta x + \gamma} \tag{24}$$

The same analysis yields a similar form for g:

$$g(y) = \frac{\alpha_0' y^2 + \beta_0' y + \gamma_0'}{\alpha' x^2 + \beta' x + \gamma'} \tag{25}$$

For $\overline{v}$ to be always positive, the coefficients (α, β, γ) in (20) must satisfy one of the following conditions:

1. $\alpha > 0$ and $4\alpha\gamma - \beta^2 > 0$.
2. $\alpha = \beta = 0$ and $\gamma > 0$.

Similarly, for $\overline{w}$ to be always positive, the coefficients $(\alpha', \beta', \gamma')$ in (21) must satisfy one of the following conditions:

1'. $\alpha' > 0$ and $4\alpha'\gamma' - \beta'^2 > 0$.

2'. $\alpha' = \beta' = 0$ and $\gamma' > 0$.

First case: 1. + 1'. In the first case, we conclude that $v = \frac{1}{Q}$ and $f = \frac{P}{Q}$ where Q is a polynomial of degree two, $Q > 0$ and P is a polynomial of degree two or less. Furthermore, $\lim_{-\infty} f = \lim_{+\infty} f = \frac{\alpha_0}{\alpha}$, regardless of whether α_0 is zero or not. This implies that f can't be invertible. Going back to (15) and plugging these expressions:

$$\begin{aligned}\log p_x(x) + \frac{1}{2}\log(\alpha x^2 + \beta x + \gamma) - \frac{1}{2}\frac{\left(\alpha_0 x^2 + \beta_0 x + \gamma_0\right)^2}{\alpha x^2 + \beta x + \gamma} - \gamma_0' x + \frac{1}{2}\gamma' x^2 + (\alpha_0 x^2 + \beta_0 x)y - \frac{1}{2}(\alpha x^2 + \beta x)y^2 \\ = \log p_y(y) + \frac{1}{2}\log(\alpha' y^2 + \beta' y + \gamma') - \frac{1}{2}\frac{\left(\alpha_0' y^2 + \beta_0' y + \gamma_0'\right)^2}{\alpha' y^2 + \beta' y + \gamma'} - \gamma_0 y + \frac{1}{2}\gamma y^2 + (\alpha_0' y^2 + \beta_0' y)x - \frac{1}{2}(\alpha' y^2 + \beta' y)x^2\end{aligned} \tag{26}$$

or again

$$A(x) - B(y) - \frac{1}{2}(\alpha - \alpha')x^2y^2 + \left(\alpha_0 - \frac{1}{2}\beta'\right)x^2y - \left(\alpha_0' - \frac{1}{2}\beta\right)xy^2 + (\beta_0 - \beta_0')xy = 0 \tag{27}$$

where

$$A(x) = \log p_x(x) + \frac{1}{2}\log(\alpha x^2 + \beta x + \gamma) - \frac{1}{2}\frac{\left(\alpha_0 x^2 + \beta_0 x + \gamma_0\right)^2}{\alpha x^2 + \beta x + \gamma} - \gamma_0' x + \frac{1}{2}\gamma' x^2 \tag{28}$$

$$B(y) = \log p_y(y) + \frac{1}{2}\log(\alpha' y^2 + \beta' y + \gamma') - \frac{1}{2}\frac{\left(\alpha_0' y^2 + \beta_0' y + \gamma_0'\right)^2}{\alpha' y^2 + \beta' y + \gamma'} - \gamma_0 y + \frac{1}{2}\gamma y^2 \tag{29}$$

By first setting $x = 0$ in equation (27), we find that $A(x) = B(0)$. Similarly, by now setting $y = 0$, we find that $B(y) = A(0)$. This in particular means that $A(x) - B(y)$ is constant, which, when plugged back in equation (27), would imply that all the monomials are zero. Finally, this would in turn imply the following:

$$\alpha = \alpha', \ \alpha_0 = -\frac{1}{2}\beta', \ \alpha_0' = -\frac{1}{2}\beta, \ \beta_0 = \beta_0' \tag{30}$$

$$\log p_x(x) = -\frac{1}{2}\gamma' x^2 + \gamma_0' x + \frac{1}{2}\frac{\left(\alpha_0 x^2 + \beta_0 x + \gamma_0\right)^2}{\alpha x^2 + \beta x + \gamma} - \frac{1}{2}\log(\alpha x^2 + \beta x + \gamma) + C \tag{31}$$

$$\log p_y(y) = -\frac{1}{2}\gamma y^2 + \gamma_0 y + \frac{1}{2}\frac{\left(\alpha_0' y^2 + \beta_0' y + \gamma_0'\right)^2}{\alpha' y^2 + \beta' y + \gamma'} - \frac{1}{2}\log(\alpha' y^2 + \beta' y + \gamma') + C \tag{32}$$

Next we need to ensure we have well-defined probability densities. From the above equations, we can check the coefficient of the quadratic term, which dominates at infinity, is $\frac{1}{2\alpha}(\alpha_0^2 - \alpha\gamma')$ for p_x. Requiring this to be negative is exactly the condition for the density family we made in Definition 1.

For p_y, we get the dominant quadratic term with the coefficient $\frac{1}{2\alpha'}(\alpha_0'^2 - \alpha'\gamma)$, and with substitutions we find the condition for its negativity as $\beta^2 < 4\alpha\gamma$ which is, again, the same as a condition in the Definition.

Second, the constant C has to be such that the probability density functions integrate to one. In fact, C can be freely chosen, but importantly, it has to be the same for both densities. As a special case, this constraint is obviously fulfilled if the densities are the same, i.e. the parameters with and without prime are the same ($\alpha = \alpha'$ etc.). We shall show below that such parameter values can be found.

In fact, we can see how the parameters of the inverse model are determined from the parameters of the true model as follows. Define

$$\delta := \gamma', \delta_0 := \gamma_0' \tag{33}$$

So we can write the above as

$$\log p_x(x) = -\frac{1}{2}\delta x^2 + \delta_0 x + \frac{1}{2}\frac{\left(\alpha_0 x^2 + \beta_0 x + \gamma_0\right)^2}{\alpha x^2 + \beta x + \gamma} - \frac{1}{2}\log(\alpha x^2 + \beta x + \gamma) + C \tag{34}$$

$$\log p_y(y) = -\frac{1}{2}\gamma y^2 + \gamma_0 y + \frac{1}{2}\frac{\left(-\beta y^2/2 + \beta_0 y + \delta_0\right)^2}{\alpha y^2 - 2\alpha_0 y + \delta} - \frac{1}{2}\log(\alpha y^2 - 2\alpha_0 y + \delta) + C \tag{35}$$

where all the parameters defining p_y are now obtained from the parameters defining p_x, f, v (which are here denoted by the parameters without prime for this specific purpose). Likewise, we see that we also get g and w using those same parameters.

Now, we show that in spite of the different constraints, a solution in this family does exist. Let us consider the case where $p_x = p_y$, which would ensure that we can normalize the densities with a common C. This can be achieved by equating corresponding constants above which only requires

$$\beta = -2\alpha_0 \tag{36}$$

$$\delta = \gamma \tag{37}$$

$$\delta_0 = \gamma_0 \tag{38}$$

which is still perfectly possible, even considering the constraints on the parameters in the Definition, which can be satisfied by simply taking non-negative α, γ, γ', and then fixing α_0 to be small enough in absolute value (which implies the same for β). Thus, a solution for the inverse direction does exist. (But note we didn't prove that it exists for data coming from any p_x, f, v in our family; we have proven unidentifiability only for some parameter values.)

Second case: 2. + 2'. In the second case, we have that v is constant. Going back to (15), multiplying by -2, plugging the solutions just obtained:

$$-2\log p_x(x) + \gamma\left(y - \frac{\alpha_0}{\gamma}x^2 - \frac{\beta_0}{\gamma}x - \frac{\gamma_0}{\gamma}\right)^2 - \log\gamma = -2\log p_y(y) + \gamma'\left(x - \frac{\alpha_0'}{\gamma'}y^2 - \frac{\beta_0'}{\gamma'}y - \frac{\gamma_0'}{\gamma'}\right)^2 - \log\gamma' \tag{39}$$

which can be expanded into, after grouping together monomials:

$$\begin{aligned} &-2\log p_x(x) + \frac{\alpha_0^2}{\gamma}x^4 + 2\frac{\alpha_0\beta_0}{\gamma}x^3 + \left(\frac{\beta_0^2 + \alpha_0\gamma_0}{\gamma} - \gamma'\right)x^2 + 2\left(\frac{\beta_0\gamma_0}{\gamma} + \gamma_0'\right)x - 2\alpha_0 x^2 y - 2\beta_0 xy + \text{const} \\ &\quad = -2\log p_y(y) + \frac{{\alpha_0'}^2}{\gamma'}y^4 + 2\frac{\alpha_0'\beta_0'}{\gamma'}y^3 + \left(\frac{{\beta_0'}^2 + \alpha_0'\gamma_0'}{\gamma'} - \gamma\right)y^2 + 2\left(\frac{\beta_0'\gamma_0'}{\gamma'} + \gamma_0\right)y - 2\alpha_0' y^2 x - 2\beta_0' xy \end{aligned} \tag{40}$$

or again

$$A(x) - B(y) - 2\alpha_0 x^2 y + 2\alpha_0' y^2 x + 2(\beta_0' - \beta_0)xy = \text{const} \tag{41}$$

where

$$A(x) = -2\log p_x(x) + \frac{\alpha_0^2}{\gamma}x^4 + 2\frac{\alpha_0\beta_0}{\gamma}x^3 + \left(\frac{\beta_0^2 + \alpha_0\gamma_0}{\gamma} - \gamma'\right)x^2 + 2\left(\frac{\beta_0\gamma_0}{\gamma} + \gamma_0'\right)x \tag{42}$$

$$B(y) = -2\log p_y(y) + \frac{{\alpha_0'}^2}{\gamma'}y^4 + 2\frac{\alpha_0'\beta_0'}{\gamma'}y^3 + \left(\frac{{\beta_0'}^2 + \alpha_0'\gamma_0'}{\gamma'} - \gamma\right)y^2 + 2\left(\frac{\beta_0'\gamma_0'}{\gamma'} + \gamma_0\right)y \tag{43}$$

By setting $y = 0$ in (41), we have that $A(x) = \text{const}$ for all x. Similarly, by setting $x = 0$, we get $B(y) = \text{const}$ for all y. We conclude that the remaining monomials must be zero. In particular, this implies that $\alpha_0 = \alpha_0' = 0$ and $\beta_0 = \beta_0'$. This in turn means that f and g are linear.

Finally, by plugging this into (42) and (43), we get:

$$\log p_x(x) = \frac{1}{2}\left(\frac{\beta_0^2}{\gamma} - \gamma'\right)x^2 + \left(\frac{\beta_0\gamma_0}{\gamma} + \gamma_0'\right)x + \text{const} \tag{44}$$

$$\log p_y(y) = \frac{1}{2}\left(\frac{{\beta_0'}^2}{\gamma'} - \gamma\right)y^2 + \left(\frac{\beta_0'\gamma_0'}{\gamma'} + \gamma_0\right)y + \text{const}' \tag{45}$$

We deduce that x and y must be Gaussian. We don't prove the normalizability of the probability density functions in detail here since it is well-known that such Gaussian, unidentifiable models exist.

Third (and fourth) case: 1. + 2'. or 2. + 1'. Since these two cases are symmetric, we will suppose that v is constant (2.) and $\overline{w}$ is a polynomial of second degree (1'.). Going back to (15) and plugging the expressions for

f, v, g, w:

$$\begin{aligned}\log p_y(y) + \frac{1}{2}\log(\alpha' y^2 + \beta' y + \gamma') - \frac{1}{2}\frac{\left(\alpha'_0 y^2 + \beta'_0 y + \gamma'_0\right)^2}{\alpha' y^2 + \beta' y + \gamma'} - \gamma_0 y + \frac{1}{2}\gamma y^2 + (\alpha'_0 y^2 + \beta'_0 y)x - \frac{1}{2}(\alpha' y^2 + \beta' y)x^2 \\ = \log p_x(x) + \frac{1}{2}\log(\gamma) - \frac{1}{2}\frac{\left(\alpha_0 x^2 + \beta_0 x + \gamma_0\right)^2}{\gamma} - \gamma'_0 x + \frac{1}{2}\gamma' x^2 + (\alpha_0 x^2 + \beta_0 x)y \end{aligned} \tag{46}$$

or again

$$A(x) - B(y) + \frac{1}{2}\alpha' x^2 y^2 + \left(\alpha_0 - \frac{1}{2}\beta'\right) x^2 y - \alpha'_0 x y^2 + (\beta_0 - \beta'_0)xy = 0 \tag{47}$$

where

$$A(x) = \log p_x(x) + \frac{1}{2}\log(\gamma) - \frac{1}{2}\frac{\left(\alpha_0 x^2 + \beta_0 x + \gamma_0\right)^2}{\gamma} - \gamma'_0 x + \frac{1}{2}\gamma' x^2 \tag{48}$$

$$B(y) = \log p_y(y) + \frac{1}{2}\log(\alpha' y^2 + \beta' y + \gamma') - \frac{1}{2}\frac{\left(\alpha'_0 y^2 + \beta'_0 y + \gamma'_0\right)^2}{\alpha' y^2 + \beta' y + \gamma'} - \gamma_0 y + \frac{1}{2}\gamma y^2 \tag{49}$$

Proceeding like above, we can deduce that $A(x) - B(y)$ is a constant, and that all the monomials in (47) are zero. In particular, $\alpha' = 0$, which contradicts 1'.: this third case is thus not possible. □

B Affine flows are not universal density approximators

Proposition 1. *Let $\mathbf{T} : \mathbb{R}^d \to \mathbb{R}^d$ be an affine autoregressive transformation. Let $\mathbf{z}$ be a standard Gaussian, and let $\mathbf{x} = \mathbf{T}(\mathbf{z})$. Then there is no parameterization of $\mathbf{T}$ such that $\mathbf{x}$ has an isotropic Gumbel distribution.*

Proof. It is enough to prove this Theorem for $d = 2$. Let $\mathbf{x} = \mathbf{T}(\mathbf{z})$. Then

$$\log p_{\mathbf{x}}(\mathbf{x}) = \log p_{\mathbf{z}}(\mathbf{T}^{-1}(\mathbf{x})) + \log |\det J_{\mathbf{T}^{-1}}(\mathbf{x})| \tag{50}$$

and

$$x_1 = e^{s_1} z_1 + t_1 \tag{51}$$

$$x_2 = e^{s_2(x_1)} z_2 + t_i(x_1) \tag{52}$$

The Jacobian log-determinant of $\mathbf{T}^{-1}$ is simply $\log |\det J_{\mathbf{T}^{-1}}(\mathbf{z})| = -s_1 - s_2(x_1)$. Note that this determinant is only a function of x_1. This is the main reason why affine autoregressive flows are not universal density approximators.

To see this, suppose that x_1 and x_2 are independent, and that each has a Gumbel distribution. Plugging this into equation (50), we get

$$-\left(x_1 + e^{-x_1}\right) - \left(x_2 + e^{-x_2}\right) = -s_1 - s_2(x_1) - (x_1 - t_1)^2 e^{-2s_1} - (x_2 - t_2(x_1))^2 e^{-2s_2(x_1)} \tag{53}$$

This equation is valid for all $(x_1, x_2) \in \mathbb{R}^2$. In particular, let $x_1 = 0$. Then for any x_2, after rearranging and grouping terms, we get

$$e^{-x_2} = \alpha x_2^2 + \beta x_2 + \gamma \tag{54}$$

This can't hold for all values of x_2, which results in a contradiction. Thus, we conclude that an affine autoregressive flow can't represent any distribution, unlike general unconstrained autoregressive flows. □

C Affine autoregressive flows are transitive

Proposition 2. *Consider 2 autoregressive transformations $\mathbf{f}$ and $\mathbf{g}$ with the same ordering π. Then their composition $\mathbf{h} = \mathbf{g} \circ \mathbf{f}$ is also an autoregressive with the same ordering π.*

Proof. Without loss of generality, assume that π is the identity. Let $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ be such that

$$\mathbf{y} = \mathbf{f}(\mathbf{z}) \tag{55}$$

$$\mathbf{x} = \mathbf{g}(\mathbf{y}) = \mathbf{g} \circ \mathbf{f}(\mathbf{z}) \tag{56}$$

Since $\mathbf{f}$ and $\mathbf{g}$ are autoregressive, we can rewrite this system using equation (4) as

$$y_i = \tau(z_i, \mathbf{y}_{<i}) \tag{57}$$

$$x_j = \tau'(y_j, \mathbf{x}_{<j}) \tag{58}$$

The transformers τ and τ' are invertible with respect to their first argument. Denoting those inverses as α and α'. Then

$$z_i = \alpha(y_i, y_{<i}) \tag{59}$$

$$y_j = \alpha'(x_j, x_{<j}) \tag{60}$$

And thus

$$z_i = \alpha(\alpha'(x_i, x_{<i}), \beta(x_{<i})) \tag{61}$$

for some function β (not necessarily invertible). Since α and α' are invertible with respect to their first argument, this means that the mapping $x_i \mapsto z_i$ in equation (61) is also invertible, and we can write:

$$x_i = \tau''(z_i, x_{<i}) \tag{62}$$

where τ'' is invertible wrt to its first argument. This proves that $\mathbf{h} = \mathbf{g} \circ \mathbf{f}$ is also an autoregressive flow. □

Proposition 3. *Consider k affine autoregressive flows $\mathbf{T}_1, \dots, \mathbf{T}_k$ of the form (6) with the same ordering π. Then their composition $\mathbf{T} = \mathbf{T}_1 \circ \cdots \circ \mathbf{T}_k$ is also an affine autoregressive flow of the form (6) with the same ordering π.*

Proof. We will suppose that $d = 2$. The proof for $d > 2$ is very similar but requires more complex notations. We will denote by z_l^j the j-th ($j = 1, 2$) output of the l-th sub-flow. Not that we can parameterize $\mathbf{T}$ or $\mathbf{T}^{-1}$ to be an affine transformation. In these notations, if $\mathbf{T}$ follows (6), then $\mathbf{z}^k = \mathbf{z}$ and $\mathbf{z}^0 = \mathbf{x}$. If instead, $\mathbf{T}^{-1}$ follows (6), then $\mathbf{z}^0 = \mathbf{z}$ and $\mathbf{z}^k = \mathbf{x}$. Each flow $l \geq 1$ has the expression:

$$z_1^l = \left(z_1^{l-1} - t_1^l\right) e^{-s_1^l} \tag{63}$$

$$z_2^l = \left(z_2^{l-1} - t_2^l(z_1^l)\right) e^{-s_2^l(z_1^l)} \tag{64}$$

First, define

$$\overline{s}_1^{l,k} = \sum_{j=l+1}^{k} s_1^j \tag{65}$$

$$\overline{t}_1^{l,k} = \sum_{j=l+1}^{k} t_1^j e^{\sum_{i=l+1}^{j-1} s_1^i} = \sum_{j=l+1}^{k} t_1^j e^{\overline{s}_1^{l,j-1}} \tag{66}$$

where all sums are zero if they have no summands. Then it is easy to show by induction using (63) that

$$z_1^l = e^{\overline{s}_1^{l,k}} z_1^k + \overline{t}_1^{l,k}, \ \forall l \leq k \tag{67}$$

and that

$$e^{\overline{s}_1^{l,j}} z_1^j + \overline{t}_1^{l,j} = e^{\overline{s}_1^{l,k}} z_1^k + \overline{t}_1^{l,k}, \ \forall l \leq \min(j,k) \tag{68}$$

Second, define

$$\overline{s}_2^k(u) = \sum_{l=1}^{k} s_2^l(e^{\overline{s}_1^{l,k}} u + \overline{t}_1^{l,k}) \tag{69}$$

$$\overline{t}_2^k(u) = \sum_{l=1}^{k} t_2^l(e^{\overline{s}_1^{l,k}} u + \overline{t}_1^{l,k}) e^{\sum_{i=1}^{l-1} s_2^i(e^{\overline{s}_1^{i,k}} u + \overline{t}_1^{i,k})} \tag{70}$$

We will show by induction on k that

$$z_2^k = \left(z_2^0 - \bar{t}_2^k(z_1^k)\right) e^{-\bar{s}_2^k(z_1^k)} \tag{71}$$

The case for $k = 1$ trivially holds. Now suppose that (71) holds for $k \geq 1$, and let's show it also holds for $k + 1$. Using (64), we can write

$$z_2^{k+1} = \left(z_2^k - t_2^{k+1}(z_1^{k+1})\right) e^{-s_2^{k+1}(z_1^{k+1})} \tag{72}$$

We need to show that

$$\bar{s}_2^{k+1}(z_1^{k+1}) = s_2^{k+1}(z_1^{k+1}) + \bar{s}_2^k(z_1^k) \tag{73}$$

$$\bar{t}_2^{k+1}(z_1^{k+1}) = t_2^{k+1}(z_1^{k+1}) e^{\bar{s}_2^k(z_1^k)} + \bar{t}_2^k(z_1^k) \tag{74}$$

This can be done using (68), the fact that $z_1^{k+1} = e^{\bar{s}_1^{k+1,k+1}} z_1^{k+1} + \bar{t}_1^{k+1,k+1}$ and the definitions of $\bar{s}_2^k$ and $\bar{t}_2^k$, which in turn allows us to conclude the induction proof.

Finally, by replacing $\mathbf{z}^0$ and $\mathbf{z}^k$ by $\mathbf{x}$ and $\mathbf{z}$ respectively, in (67) and (71), we have

$$x_1 = e^{\bar{s}_1^{0,k}} z_1 + \bar{t}_1^{0,k} \tag{75}$$

$$x_2 = e^{\bar{s}_2^k(x_1)} z_2 + \bar{t}_2^k(x_1) \tag{76}$$

which proves the transitivity of affine autoregressive flows. □

D Universality of the causal function

Proposition 4. *Consider k affine autoregressive flows $\mathbf{T}_1, \dots, \mathbf{T}_k$, and let $\mathbf{T} = \mathbf{T}_1 \circ \dots \circ \mathbf{T}_k$. Denote by t_j^l and s_j^l the coefficients of the l-th sub-flow $\mathbf{T}_l$, and by $\bar{t}_j^k$ and $\bar{s}_j^k$ those of $\mathbf{T}$. Suppose that all of the s_j^l and t_j^l are feed-forward neural networks that have universal approximation capability (assuming all technical conditions hold). Then $\bar{t}_j^k$ and $\bar{s}_j^k$ also have universal approximation capability.*

Proof. We will suppose for the proof that $d = 2$. The proof for $d > 2$ is similar. According to Proposition 3, $\mathbf{T}$ is also an affine autoregressive flow, and $\bar{t}_2^k$ and $\bar{s}_2^k$ have the following expressions:

$$\bar{s}_2^k(u) = \sum_{l=1}^k s_2^l\left(e^{\bar{s}_1^{l,k}} u + \bar{t}_1^{l,k}\right) \tag{77}$$

$$\bar{t}_2^k(u) = \sum_{l=1}^k t_2^l\left(e^{\bar{s}_1^{l,k}} u + \bar{t}_1^{l,k}\right) e^{\sum_{i=1}^{l-1} s_2^i\left(e^{\bar{s}_1^{i,k}} u + \bar{t}_1^{i,k}\right)} \tag{78}$$

where $\bar{t}_1^{l,k}$ and $\bar{s}_1^{l,k}$ are defined in equations (66) and (65) respectively.

On the one hand, translating and scaling the argument u of $\bar{s}_2^k$ by $\bar{t}_1^{l,k}$ and $\bar{s}_1^{l,k}$ only changes the bias and the slope of the input layer of each of the s_2^l, $l = 1, \dots, k$. Thus, one can interpret equation (77) as the output of an additional final layer of the neural network whose outputs are the s_2^l functions. The number of flows k in this case increases the width of this final layer. Using the classical result of the universal approximation theorem of feed-forward networks with arbitrary width (Hornik, 1991), we conclude that $\bar{s}_2^k$ also satisfies such properties.

Interestingly, note that this results holds even if each of the s_j^l function is simply an affine function followed by a nonlinearity (*i.e.* a 1-hidden layer feed-forward network).

On the other hand, since each of the t_2^l have universal approximation capability, each can in particular approximate a function of the form $u \mapsto f_l(u) e^{\sum_{i=1}^{l-1} s_2^i\left(e^{\bar{s}_1^{i,k}} u + \bar{t}_1^{i,k}\right)}$, where f_l is a simple affine function followed by a nonlineariy σ (*i.e.* a 1-hidden layer feed-forward network). Thus, $\bar{t}_2^k$ can approximate a function of the form $\sum_{l=1}^k f_l$, which, by the same argument used above, will have universal approximation capability (Hornik, 1991). □

E Algorithms for causal inference

E.1 Interventions

As discussed in Section 4.1, the intervention $\mathrm{do}(x_i = \alpha)$ breaks the links from $x_{<\pi(i)}$ to x_i and sets a point mass on z_i. Computing the value of $\mathbf{x}_{j\neq i}$ requires sampling from $\prod_{j\neq i} p_{z_j}$ then propagating sequentially through the flow. This avoids having to invert the flow and compute $z_i = \tau_i^{-1}(x_i, \mathbf{x}_{<\pi(i)})$. However, in the case of affine autoregressive flows, tau^{-1} is readily available, and can be used to make the above algorithm parallelizable. In fact, we can compute $z_i = \tau_i^{-1}(x_i, \mathbf{x}_{<\pi(i)})$, sample $\mathbf{z}_{j\neq i}$, then propagate the concatenated $\mathbf{z}$ forward through the flow to obtain $\mathbf{x}^{\mathrm{do}(x_i=\alpha)}$. Note that the value of $\mathbf{x}_{<\pi(i)}$ is required to infer z_i, which will break the parallelism. But since the same value is used to parametrized τ_i and τ_i^{-1}, any value $\mathbf{v}$ can be used as long as $\tau_i(\tau_i^{-1}(\alpha, \mathbf{v}), \mathbf{v}) = \alpha$. In our implementation, we chose $\mathbf{v} = 0$. The sequential and parallel implementation are summarized by Algorithms 1 and 2 respectively.

Algorithm 1 Generate samples from an interventional distribution (sequential)

```
Input: interventional variable $x_i$, intervention value $\alpha$, number of samples $S$
for $s = 1$ to $S$ do
    sample $\mathbf{z}(s)$ from flow base distribution (the value of $z_i$ can be discarded)
    set $x_i(s) = \alpha$
    for $j = \pi^{-1}(1)$ to $\pi^{-1}(d)$; $j \neq i$ do
        compute observation $x_j(s) = \tau_j(z_j(s), \mathbf{x}_{<\pi(j)}(s))$
    end for
end for
Return: interventional sample $\mathbf{X} = \{\mathbf{x}(s) : s = 1, \dots, S\}$
```

Algorithm 2 Generate samples from an interventional distribution (parallel)

```
Input: interventional variable $x_i$, intervention value $\alpha$, number of samples $S$
for $s = 1$ to $S$ do
    sample $\mathbf{z}(s)$ from flow base distribution (the value of $z_i$ can be discarded)
    set $z_i(s) = \tau^{-1}(\alpha, \mathbf{0})$
    compute $\mathbf{x}(s) = \mathbf{T}(\mathbf{z}(s))$
end for
Return: interventional sample $\mathbf{X} = \{\mathbf{x}(s) : s = 1, \dots, S\}$
```

E.2 Counterfactuals

The process of obtaining counterfactual predictions is described in Pearl (2009a) as consisting of three steps:

1. **Abduction**: given an observation $\mathbf{x}^{obs}$, infer the conditional distribution/values over latent variables $\mathbf{z}^{obs}$. In the context of an autoregressive flow model this is obtained as $\mathbf{z}^{obs} = \mathbf{T}^{-1}(\mathbf{x}^{obs})$.

2. **Action**: substitute the values of $\mathbf{z}^{obs}$ with the values based on the counterfactual query, $\mathbf{x}_{x_j \leftarrow \alpha}$. More concretely, for a counterfactual, $\mathbf{x}_{x_j \leftarrow \alpha}$, we replace the structural equations for x_j with $x_j = \alpha$ and adjust the inferred value of latent z_j^{obs} accordingly.

3. **Prediction**: compute the implied distribution over $\mathbf{x}$ by propagating latent variables, $\mathbf{z}^{obs}$, through the structural equation models.

This is summarized by Algorithm 3.

F Experimental details

F.1 Architectures and hyperparameters

The optimization was done using Adam, with learning rate $\mathtt{lr} = 0.001$, $\boldsymbol{\beta} = (0.9, 0.999)$, along with a scheduler that reduces the learning rate by a factor of 0.1 on plateaux. All flows use an isotropic Laplace distribution as a prior. The different architectures and hyperparameter used for the experiments are as follows:

- **Causal discovery simulations:** The flow $\mathbf{T}$ is a composition of 2 sub-flows $\mathbf{T}_1$ and $\mathbf{T}_2$. For each of the $\mathbf{T}_l$, both s_j and t_j are multi-layer perceptrons (MLPs), with 1 hidden layer and 10 hidden units. Each direction was trained for 200 epochs, with a mini-batch of 128 data points. The same architecture was used for all panels of Figure 1.
- **Cause-effect pairs:** The flow $\mathbf{T}$ is a composition of 4 sub-flows $\mathbf{T}_1, \cdots, \mathbf{T}_4$. For each of the $\mathbf{T}_l$, both s_j and t_j are MLPs, with either 1 or 3 hidden layers, each with 5 hidden units. For each direction, we train two different flows (with 1 or 3 hidden layers), and select the flow that yields higher test likelihood. Each direction was trained for 750 epochs, with a mini-batch of 128 data points. For each pair, 80% of the data points were used for training, and the remaining 20% to evaluate the likelihood. The same architecture was used to classify all the pairs.
- **EEG arrow of time:** The flow $\mathbf{T}$ is a composition of 4 sub-flows $\mathbf{T}_1, \cdots, \mathbf{T}_4$. For each of the $\mathbf{T}_l$, both s_j and t_j are MLPs, with 4 hidden layers, each with 10 hidden units. Each direction was trained for 400 epochs, with a mini-batch of 32 data points. For each channel, 80% of the data points were used for training, and the remaining 20% to evaluate the likelihood. The same architecture was used to classify all the channels.
- **Interventions on simulated data:** The flow $\mathbf{T}$ is a composition of 5 sub-flows $\mathbf{T}_1, \cdots, \mathbf{T}_5$. For each of the $\mathbf{T}_l$, both s_j and t_j are MLPs, with 1 hidden layers, each with 10 hidden units. We train the flow, conditioned on the causal ordering, to fit the correct SEM. Training was done for 750 epochs, with a mini-batch of 32 data points.
- **Interventions on es-fMRI data:** The flow $\mathbf{T}$ is a composition of 5 sub-flows $\mathbf{T}_1, \cdots, \mathbf{T}_5$. For each of the $\mathbf{T}_l$, both s_j and t_j are MLPs, with a single hidden layer consisting of 2 hidden units. In order to obtain interventional predictions, a CAREFL model was first trained using resting-state fMRI data conditioned upon the causal ordering. Since we did not seek to infer the causal structure, 100% of the training data was employed (this is in contrast to causal discovery experiments which only trained models on 80% of the data).

F.2 Exploring flow architectures

As discussed in Section C, stacking multiple autoregressive flows on top of each other is equivalent to using a single autoregressive flow with a wide hidden layer. To explore this aspect, we run multiple experiments where each flow is an MLP with one hidden layer and a LeakyReLU activation, in which we vary the width of the hidden layer and the number of stacked flows. We observed empirically that stacking multiple layers in the flows lead to empirical improvements, as reported by Figure 6.

F.3 Preprocessing of EEG data

The openly available EEG data from Dornhege et al. (2004) contains recordings for 5 healthy subjects. For each subject, the data has been sampled at 100Mhz and 1000Mhz. For our experiments, we considered subject number

Algorithm 3 Answer a counterfactual query

```
**Input:** observed data $\mathbf{x}^{obs}$, counterfactual variable $x_j$ and value $\alpha$
  1. **Abduction**: infer $\mathbf{z}^{obs} = \mathbf{T}^{-1}(\mathbf{x}^{obs})$
  2. **Action**: $(a)$ set $z^{obs}_{j,x_j\leftarrow\alpha} = \tau_j^{-1}(\alpha, \mathbf{x}^{obs}_{<\pi(j)})$
               $(b)$ set $z^{obs}_{i,x_j\leftarrow\alpha} = z_i^{obs}$ for $i \neq j$
  3. **Prediction**: pass $\mathbf{z}^{obs}_{x_j\leftarrow\alpha}$ forward through the flow $\mathbf{T}$
**Return:** $\mathbf{x}_{x_j\leftarrow\alpha} = \mathbf{T}(\mathbf{z}^{obs}_{x_j\leftarrow\alpha})$
```

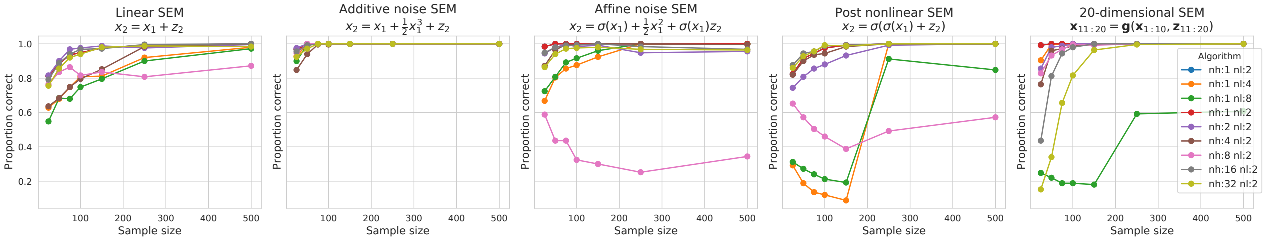

Figure 6: Impact of changing the width versus the depth of the normalizing flow in CAREFL

3, and used the data sampled at 1000Mhz. In particular, we only considered $n = 150$ and $n = 500$ time points. Each of the 118 EEG channels was then reversed with probability 0.5.

The task is to properly infer the arrow of time for each of the 118 EEG, considered separately. We transform a univariate timeseries $(x_t)_{t\in[\![1,n]\!]}$ corresponding to 1 channel into bivariate causal data by shifting it by a lag parameter l, to obtain data of the form $(x_t, x_{t+l})_{t\in[\![1,n-l]\!]}$. For the results plotted in Figure 3, we used three values of lag for ANM, RECI, the linear LR and CAREFL-NS: $l \in \{1, 2, 3\}$, which we then combined into one dataset. For CAREFL, we used only two values of lag: $l \in \{1, 2\}$.

F.4 Preprocessing of functional MRI data

Results included in this manuscript come from preprocessing performed using `FMRIPREP` (Esteban et al., 2019), a `Nipype` based tool (Gorgolewski et al., 2011). Each T1w (T1-weighted) volume was corrected for INU (intensity non-uniformity) using `N4BiasFieldCorrection v2.1.0` and skull-stripped using `antsBrainExtraction.sh v2.1.0` (using the OASIS template). Brain surfaces were reconstructed using recon-all from `FreeSurfer v6.0.1`, and the brain mask estimated previously was refined with a custom variation of the method to reconcile `ANTs`-derived and `FreeSurfer`-derived segmentations of the cortical grey-matter of Mindboggle. Spatial normalization to the ICBM 152 Non-linear Asymmetrical template version 2009c was performed through non-linear registration with the `antsRegistration` tool of `ANTs v2.1.0`, using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using `fast`.

Functional data was slice time corrected using `3dTshift` from `AFNI v16.2.07` and motion corrected using `mcflirt`. This was followed by co-registration to the corresponding T1w using boundary-based registration with six degrees of freedom, using `bbregister` (`FreeSurfer v6.0.1`). Motion correcting transformations, BOLD-to-T1w transformation and T1w-to-template (MNI) warp were concatenated and applied in a single step using `antsApplyTransforms` using Lanczos interpolation.

Regional time series were subsequently calculated from the processed FMRI data (transformed into MNI space) using `NiLearn` (Abraham et al., 2014) and the Harvard-Atlas probabilistic atlas, with regions thresholded at 25% probability and binarised. Given the regional location of intracortical stimulation in the subjects, FMRI time-series from the Cingulate gyrus and Heschl's gyrus were selected for analysis.

We note that each patient received surgery and stimulation in different locations, as determined by their diagnosis and clinical criteria. As such, the two regions studied were selected so as to include as many subjects as possible in our experiments. Moreover, the Cingulate gyrus is a region associated with cognitive processes such as saliency and emotional processing (Vogt, 2019) whereas Heschl's gyrus covers primary auditory cortex, associated with early cortical processing of auditory information; as such connectivity between the regions captures the interaction between a higher-order heteromodal region and a unimodal sensoryl region.