Supplementary Information

Appendix A: Elements of group theory

Sets equipped with a group structure have attracted interest from the machine learning community because they can model the data structure of complex domains (Fukumizu et al., 2009). We will introduce concisely the concepts and results of group theory necessary to this paper. The authors can refer for example to (Tung, 1985; Wijsman, 1990; Eaton, 1989) for more details.

Definition 4 (Group) *A set $\mathcal{G}$ is said to form a group if there is an operation '*', called group multiplication, such that:*

1. *For any $a, b \in \mathcal{G}$, $a * b \in \mathcal{G}$.*
2. *The operation is associative: $a * (b * c) = (a * b) * c$, for all $a, b, c \in \mathcal{G}$,*
3. *There is one identity element $e \in \mathcal{G}$ such that, $g * e = e$ for all $g \in \mathcal{G}$,*
4. *Each $g \in \mathcal{G}$ has an inverse $g^{-1} \in \mathcal{G}$ such that, $g * g^{-1} = e$.*

A subset of $\mathcal{G}$ is called a subgroup if it is a group under the same multiplication operation.

The following elementary properties are a direct consequence of the above definition: $e^{-1} = e$, $g^{-1} * g = e$, $e * g = g$, for all $g \in \mathcal{G}$.

Among others, classical groups of interest in this paper are the permutations group $\mathbb{S}(n)$ and the *general linear group* $GL(n)$ of all real nonsingular $n \times n$ matrices equipped with matrix multiplication. The matrix representations of the real orthogonal group $O(n)$ of isometries and of the real special orthogonal group $SO(n)$ of rotations are subgroups of $GL(n)$. As in these two examples, many groups can be considered as functions *acting* on an input space:

Definition 5 (Action) *Let $\mathcal{G}$ be a group and $\mathcal{X}$ a space. An action of $\mathcal{G}$ on $\mathcal{X}$ to the left is a function $a : \mathcal{G} \times \mathcal{X} \to \mathcal{X}, (g, x) \mapsto g.x$ such that:*

1. *$e.x = x$, for all $x \in \mathcal{X}$*
2. *$g_2.(g_1.x) = (g_2 * g_1).x$, for all $g_1, g_2 \in \mathcal{G}$, $x \in \mathcal{X}$*

If $g*x = x$, x is called a fixed point of g. We will call the subgroup of elements fixing x, $\mathcal{G}_x = \{g \in \mathcal{G}, g*x = x\}$, the isotropy subgroup or stabilizer of x in $\mathcal{G}$. $\mathcal{G}$ is said to act freely if $gx \neq x$, for all $g \in \mathcal{G} \setminus \{e\}$ and $x \in \mathcal{X}$.

Due to the properties of group actions, associativity rules can be applied to all group actions and group multiplications of a given expression, such that we can do not need to put any symbol for binary operations between group/space elements. For example, we will thus simply denote $g_1.((g_2 * g_3).x)$ by $g_1 g_2 g_3 x$.

We will always consider group actions to the left in this paper, such that we will simply call them group action. It is easy to show that $\mathbb{S}(n)$ and its subgroups act on the set $\{1, .., n\}$ by permuting its elements, as well as on n-tuples from arbitrary sets. $GL(n)$ and its subgroups act as linear functions on the vector space $\mathbb{R}^n$.

Definition 6 (Topological group) *A locally compact Hausdorff topological group is a group equipped with a locally compact Hausdorff topology such that:*

- *$\mathcal{G} \to \mathcal{G} : x \mapsto x^{-1}$ is continuous,*
- *$\mathcal{G} \times \mathcal{G} \to \mathcal{G} : (x, y) \mapsto x.y$ is continuous (using the product topology).*

The σ-algebra generated by all open sets of G is called the Borel algebra of $\mathcal{G}$.

Definition 7 (Invariant measure) *Let $\mathcal{G}$ be a topological group according to definition 6. Let $K(\mathcal{G})$ be the set of continuous real valued functions with compact support on $\mathcal{G}$. A radon measure μ defined on Borel subsets is left invariant if for all $f \in K(\mathcal{G})$ and $g \in \mathcal{G}$*

$$\int_G f(g^{-1}x)d\mu(x) = \int_G f(x)d\mu(x)$$

Such a measure is called a Haar measure.

A key result regarding topological groups is the existence and uniqueness up to a positive constant of the Haar measure (Eaton, 1989). Whenever $\mathcal{G}$ is compact, the Haar measures are finite and we will denote $\mu_{\mathcal{G}}$ the unique Haar measure such that $\mu_{\mathcal{G}}(\mathcal{G}) = 1$, defining an invariant probability measure on the group.

Appendix B: proofs of main text results

Proposition 2

We will use $\mathbb{E}_U$ to denote the expectation when U is drawn from the distribution on $SO(n)$ (or sometimes just $\mathbb{E}$ when it does not lead to confusion). Since A is symmetric then we can decompose it as $A = V^T DV$, with D diagonal and $V \in SO(n)$. Then

$$\mathbb{E}_U \operatorname{tr}\left(U^T AUB\right) = \mathbb{E}_{VU} \operatorname{tr}\left((VU)^T DVUB\right) = \mathbb{E}_U \operatorname{tr}\left(U^T DUB\right),$$

where we substituted VU for U in the expectation and used the translation invariance of the Haar measure. As a consequence,

$$\mathbb{E} \operatorname{tr}\left(U^T AUB\right) = \mathbb{E} \operatorname{tr}\left(DUBU^T\right) = \sum d_{kk} \mathbb{E}\left(UBU^T\right)_{kk}$$

by cyclic invariance and linearity of the trace. We claim that the values of the diagonal elements of $\mathbb{E}\left(UBU^T\right)$ are all equal. Indeed, let $P_{i,k} \in SO(n)$ be the matrix permuting coordinates i and k. Then

$$\mathbb{E}\left(UBU^T\right)_{kk} = \left(P_{i,k}\left(\mathbb{E}\left(UBU^T\right)\right)P_{i,k}^T\right)_{ii} = \mathbb{E}\left(P_{i,k}UB\left(P_{i,k}U\right)^T\right)_{ii} = \mathbb{E}\left(UBU^T\right)_{ii}$$

again by substitution the translation invariance of the Haar measure. Therefore, for all k,

$$\mathbb{E}\left(UBU^T\right)_{kk} = \tfrac{1}{n}\mathbb{E} \operatorname{tr}\left(UBU^T\right) = \tfrac{1}{n}\mathbb{E} \operatorname{tr}\left(U^T UB\right) = \tfrac{1}{n} \operatorname{tr}(B)$$

since U is orthogonal. Finally we get

$$\mathbb{E}_U \operatorname{tr}\left(U^T AUB\right) = \tfrac{1}{n} \operatorname{tr}(B) \sum d_{kk} = \tfrac{1}{n} \operatorname{tr}(B) \operatorname{tr}(A).$$

Equation 3

We first observe that for all x in $\mathcal{A}$, and all $g \in \mathcal{G}$

$$\langle C\rangle_{m,gx} = \mathbb{E}_{h\sim\mu_{\mathcal{G}}} C(mhgx) = \mathbb{E}_{h'\sim\mu_{\mathcal{G}}} C(mh'x) = \langle C\rangle_{m,x}$$

due to the invariance of the Haar measure $\mu_{\mathcal{G}}$ (see appendix A). Which means that the EGC is constant on orbits of the form $\mathcal{G}x$, for all x. Thus

$$\mathbb{E}_x\left[\frac{C(mx)}{\langle C\rangle_{m,x}}\right] = \mathbb{E}_{\tilde{x}}\mathbb{E}_g\left[\frac{C(mg\tilde{x})}{\langle C\rangle_{m,g\tilde{x}}}\right] = \mathbb{E}_{\tilde{x}}\frac{\mathbb{E}_g C(mg\tilde{x})}{\langle C\rangle_{m,\tilde{x}}} = 1\,. \tag{16}$$

Indentifiability of LiNGAM using thrid order cumulants

Assume the cause X is a real random variable with zero mean such that $\mathbb{E}X^3 \neq 0$, and Y is generated from X through

$$X \mapsto Y := \alpha X + \epsilon\,, \tag{17}$$

where $\alpha \in \mathbb{R}^*$ and ϵ is a zero mean i.i.d noise random variable independent of X.

Asymptotically, we clearly have (13) satisfied and all required parameters can be estimated consistently by least squares. Let $\sigma_X^2 = \mathbb{E}[X^2]$ and $\sigma_\epsilon^2 = \mathbb{E}[\epsilon^2]$. In the backward direction, the least square solution is

$$X = \beta Y + \eta$$

with

$$\beta = \frac{\alpha\sigma_X^2}{\alpha^2\sigma_X^2 + \sigma_\epsilon^2}.$$

We get for the backward additive noise the expression

$$\eta = (1 - \alpha\beta)X - \beta\epsilon = \frac{\sigma_\epsilon^2}{\alpha^2\sigma_X^2 + \sigma_\epsilon^2}X - \frac{\alpha\sigma_X^2}{\alpha^2\sigma_X^2 + \sigma_\epsilon^2}\epsilon.$$

By contradiction, assume the model is not identifiable, then (13) must also be satisfied for the backward model. The asymptotic expression is

$$\kappa_3(\beta Y + \eta) = \kappa_3(\beta Y) + \kappa_3(\eta).$$

where κ_3 denotes the population third order centered cumulant. Which implies

$$\frac{\alpha^2\sigma_\epsilon^4\sigma_X^2}{\left(\alpha^2\sigma_X^2 + \sigma_\epsilon^2\right)^3}\mathbb{E}X^3 = -\frac{\alpha^4\sigma_\epsilon^2\sigma_X^4}{\left(\alpha^2\sigma_X^2 + \sigma_\epsilon^2\right)^3}\mathbb{E}X^3$$

whenever $\alpha \neq 0$ and $\sigma_\epsilon^2 \neq 0$ this is impossible, so the true causal direction can be identified.

Proposition 3

Proof. Decomposing $\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top$ using $\boldsymbol{X} = \boldsymbol{M} + \boldsymbol{V}$

$$\begin{aligned}
\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top &= \left(\boldsymbol{V}\boldsymbol{V}^\top + \boldsymbol{V}\boldsymbol{M}^\top + \boldsymbol{M}\boldsymbol{V}^\top + \boldsymbol{M}\boldsymbol{M}^\top\right)^2 \\
&= \boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{V}^\top + \boldsymbol{V}\boldsymbol{M}^\top\boldsymbol{V}\boldsymbol{M}^\top + \boldsymbol{M}\boldsymbol{V}^\top\boldsymbol{M}\boldsymbol{V}^\top + \boldsymbol{M}\boldsymbol{M}^\top\boldsymbol{M}\boldsymbol{M}^\top \\
&\quad + \boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{M}^\top + \boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{M}\boldsymbol{V}^\top + \boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{M}\boldsymbol{M}^\top \\
&\quad + \boldsymbol{V}\boldsymbol{M}^\top\boldsymbol{V}\boldsymbol{V}^\top + \boldsymbol{V}\boldsymbol{M}^\top\boldsymbol{M}\boldsymbol{V}^\top + \boldsymbol{V}\boldsymbol{M}^\top\boldsymbol{M}\boldsymbol{M}^\top \\
&\quad + \boldsymbol{M}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{V}^\top + \boldsymbol{M}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{M}^\top + \boldsymbol{M}\boldsymbol{V}^\top\boldsymbol{M}\boldsymbol{M}^\top \\
&\quad + \boldsymbol{M}\boldsymbol{M}^\top\boldsymbol{V}\boldsymbol{V}^\top + \boldsymbol{M}\boldsymbol{M}^\top\boldsymbol{V}\boldsymbol{M}^\top + \boldsymbol{M}\boldsymbol{M}^\top\boldsymbol{M}\boldsymbol{V}^\top
\end{aligned}$$

Taking the expectation and the trace and using $\operatorname{tr}\left[AB^\top CD^\top\right] = \operatorname{tr}\left[BA^\top DC^\top\right]\operatorname{tr} = \left[CD^\top AB^\top\right] = \operatorname{tr}\left[DC^\top BA^\top\right]$, we get for the contrast

$$\begin{aligned}
\mathbb{E}\operatorname{tr}\left(\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top\right) &= \mathbb{E}_z\mathbb{E}_{\boldsymbol{X}|z}\operatorname{tr}\left(\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top\right) \\
&= \sum \pi_k\left(\|\boldsymbol{\mu}_k\|^4 + \mathbb{E}_{\boldsymbol{V}|z}\operatorname{tr}\left[\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{V}^\top\right] + 4\boldsymbol{\mu}_k^\top\boldsymbol{\Sigma}_k\boldsymbol{\mu}_k + 2\|\boldsymbol{\mu}_k\|^2\operatorname{tr}\left[\boldsymbol{\Sigma}_k\right]\right)
\end{aligned}$$

We can notice that all terms but one are influenced by introducing a generic transformation $M \mapsto UM$ with $U \in SO(p)$, and the final result is obtained using proposition 2. □

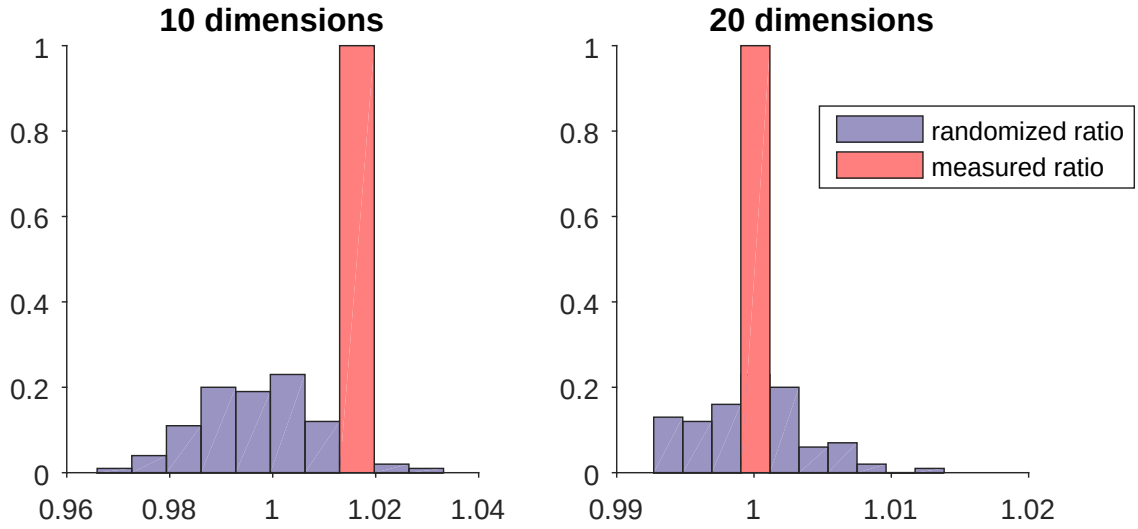

Figure 7: Normalized histograms of the generic ratio for the MNIST dataset.