WHAT IS IDENTIFIED IN CAUSAL REPRESENTATION LEARNING? EXACT ANSWERS FOR THE NORMALIZED iVAE FROM REALIZED SYMMETRIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal representation learning shows that a latent representation can be recovered only up to a symmetry. This leaves a question unanswered: which claims about hidden causes are fixed by observed data? We introduce a framework that answers it by characterizing functions of latent variables that are determined by observations and unchanged by symmetries. We first show why normalization is essential: a class that admits monotone changes of every latent coordinate identifies only constant functions. For the normalized iVAE with scalar statistics and a fixed latent parameterization, we build changes of variables that realize every range-preserving permutation and sign change. This turns the classical iVAE inclusion into an exact statement within each set of models producing the same observations: the bounded functions unaffected by those permutations and sign changes are identified. We show how these statements fit together across model classes. A rigid class preserves content, while a class allowing monotone changes in every coordinate reduces it to constants. Finally, we show how contrastive content and style separation fits the same interface and sketch how CITRIS fits the picture. Experiments with fitted iVAEs show that the symmetry is real in practice, that identified functions are the stable ones across retraining, and that a stronger normalization restores individual coordinates.

1 INTRODUCTION

Identifiability theorems in causal representation learning promise that a latent representation can be recovered from data "up to" a group of symmetries: permutations of coordinates, sign changes, monotone rescalings (Khemakhem et al., 2020; Hyvärinen et al., 2019; von Kügelgen et al., 2021; Lippe et al., 2022; Ahuja et al., 2023). The promise is a guarantee about the representation. It is not yet a guarantee about anything a practitioner computes from it. Suppose two runs of the same model on the same data return latents $\hat{z}$ and $\hat{z}'$ that differ by swapping the two coordinates and flipping the sign of one. Both are correct according to the theorem. A downstream quantity such as "the mean of coordinate one" now takes two different values, and the data cannot say which is right. The question the theorem leaves open is: *which statements about the hidden variables are fixed by the data?*

A two-coordinate example. Take a model with two latent coordinates whose fitted statistics (T_1, T_2) are recoverable up to permutation and sign flips (the group B_2 of eight signed permutations). Table 1 lists candidate quantities and whether they are determined. The "No" rows are theorems, not gaps in knowledge: under this symmetry no amount of data pins those quantities down. A pipeline that consumes per-coordinate quantities, such as a per-coordinate fairness constraint or a per-coordinate disentanglement score, is consuming a quantity that depends on an arbitrary choice of representative.

Answering the question in general requires two things. First, a definition of the identified content that does not depend on picking a representative: the bounded functions of the latents that are both determined by the observed law and unchanged by every symmetry of the model class. Second, an argument in two directions. The "up to" theorem is an upper bound on ambiguity: if two solutions disagree, they differ by *some* symmetry, so symmetric functions are identified. It does not show that nothing else is: if the sign flip were never realized by an admissible model, T_1 itself would be

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Which quantities are determined by data when (T_1, T_2) is known up to permutation and sign flips. For an unbounded quantity, "Yes" means that every bounded function of it is determined.

Candidate quantity	Determined?	Why
T_1	No	a swap turns it into T_2
$\lvert T_1\rvert$	No	sign-proof but not swap-proof
$T_1 + T_2$	No	a sign flip changes it
sorted (T_1, T_2)	No	swap-proof, but a sign flip reorders the pair
sorted $(\lvert T_1\rvert, \lvert T_2\rvert)$	Yes	a function of (T_1^2, T_2^2)
$e^{-(T_1^2+T_2^2)}$, $\tanh(T_1^2 T_2^2)$	Yes	bounded functions of symmetric functions of squares

identified after all. Closing the gap means constructing, for every symmetry, an admissible model that realizes it, and that construction is the technical heart of the paper.

Contributions.

1. *A framework and a two-ingredient test* (Section 2). We define the identified content V^B of an observation model and a model class, and prove that V^B equals the invariants of a group G whenever (i) observationally equivalent models differ by an element of G, and (ii) every element of G is realized inside the class. Ingredient (i) is what existing identifiability theorems provide; ingredient (ii) is the new work per model family.
2. *Normalization is necessary* (Section 3). In a class that admits coordinatewise monotone changes of the latents, those changes act transitively and V^B contains only constants.
3. *Exact identifiability for the normalized iVAE* (Section 4). Under explicit assumptions, we construct, for every signed permutation, a change of latent variables that stays inside the model class and preserves the observed law. Within each set of models producing the same observations, the identified content is exactly the bounded functions invariant under the signed permutations that preserve the range of the statistics. We give the derivation of the signed-permutation group in full, show by a Gaussian counterexample which assumption it rests on, and show that on bounded or one-sided ranges that assumption is a theorem rather than a hypothesis.
4. *Getting coordinates back* (Section 5). A skewness normalization with distinct margins selects one representative per orbit, so every coordinate becomes identified. We give a finite-sample guarantee for the data-driven version and a certificate that abstains when the data cannot decide.
5. *The global picture and evidence* (Sections 6 and 7). Class-local exactness lifts to a global statement on a rigid class and collapses on a fully monotone class. Fitted iVAEs show the symmetry scrambling coordinates across retraining, identified functions two to three times more stable than non-identified ones, and the skewness normalization turning coordinate identity from a coin flip into agreement in eight of eight restarts.

Scope. The framework and the two-ingredient test are general. The exact-equality theorems are about the normalized iVAE class under stated assumptions, not about causal representation learning at large, and we mark where each assumption binds. Exact identifiability is a foundational step: it tells you which downstream quantities are well defined before you estimate them.

2 Identified content and the two-ingredient test

Instances and equivalence. A latent variable Z lives in an open set $\mathcal{Z} \subseteq \mathbb{R}^m$ with law P_Z. An *instance* $I = (P_Z, \kappa, f)$ consists of the latent law, an auxiliary channel κ that generates a side variable S from Z, and a decoder f that generates the observation $X = f(Z)$. An *observation model* Π fixes how (X, S) is generated from an instance, and a *model class* $\mathcal{M}$ is a set of instances sharing structural constraints. Two instances are *observationally equivalent*, $I \sim_\Pi I'$, if they induce the same law of (X, S); the equivalence class of I is written $[I]_{\sim_\Pi}$. A measurable bijection ψ of $\mathcal{Z}$ acts on instances by relabeling the latents, $\psi \cdot I := (\psi_\# P_Z,\ \kappa \circ \psi^{-1},\ f \circ \psi^{-1})$, which never changes the law of (X, S). Throughout, every latent law in a class has a strictly positive density on $\mathcal{Z}$ with respect to Lebesgue measure, so "null set" and "almost surely" (a.s.) refer to Lebesgue measure and mean

the same thing for every instance of the class; $L^\infty(\mathcal{Z})$ denotes the bounded Borel functions on $\mathcal{Z}$ modulo null sets.

Definition 1 (Identified content)**.** *The* expectation-identified *functions are* $V^A(\Pi, \mathcal{M}) := \{\varphi \in L^\infty(\mathcal{Z}) : \int \varphi \, dP_Z^I = \int \varphi \, dP_Z^{I'}$ *whenever* $I \sim_\Pi I'$ *in* $\mathcal{M}\}$*. The* symmetries *of the class are the measurable bijections* ψ *of* $\mathcal{Z}$ *that map* $\mathcal{M}$ *onto itself and preserve the observed law of every instance, written* $\mathrm{Aut}(\mathcal{Z})_{\Pi,\mathcal{M}}$*. The* identified content *is*

$$V^B(\Pi, \mathcal{M}) := \{\varphi \in V^A(\Pi, \mathcal{M}) : \varphi \circ \psi = \varphi \text{ a.s. for all } \psi \in \mathrm{Aut}(\mathcal{Z})_{\Pi,\mathcal{M}}\}.$$

In words: φ is identified if its expectation is fixed by the observed law and its value does not change under any relabeling of the latents that the class cannot distinguish. Describing what is identified as the functions of a maximal invariant of a group action is classical in the theory of invariant statistical models (Lehmann & Romano, 2005, Ch. 6); see also Eaton (1989). What is different here is that the group is not given in advance: it has to be determined from the observation model and the class, and determining it is the work of this paper. Both conditions are needed. The first alone would count every bounded function of T_1 as identified in a class with only one instance per equivalence class; the second alone would ignore the data.

Class-local content. A reference chart T^I attached to one instance describes the identified content of *its* equivalence class $C := [I]_{\sim_\Pi} \cap \mathcal{M}$; a different class has a different chart. Our exact statements are about $V^B(\Pi, C)$ (for the iVAE, about the chart-pinned subclass C_I of Section 4); Section 6 treats the global $V^B(\Pi, \mathcal{M})$ through its decomposition over classes.

The test. Fix a class C and a chart $T : \mathcal{Z} \to Y$, a measurable injection with measurable inverse on $T(\mathcal{Z})$, and let a group G act measurably on Y. Write "a.e. on Y" for almost everywhere with respect to the law of $T(Z)$ under the reference instance, and assume every $g \in G$ maps null sets of Y to null sets.

Theorem 2 (Two-ingredient exactness test)**.** *Suppose (i)* witnesses*: whenever* $J \sim_\Pi J'$ *in* C *there is a measurable bijection* ψ *with* $J' = \psi \cdot J$*, and every measurable bijection* ψ *with* $J' = \psi \cdot J$ *satisfies* $T \circ \psi = g_\psi \circ T$ *a.s. for some* $g_\psi \in G$ *(when decoders are injective the relabeling is unique,* $\psi = f_{J'}^{-1} \circ f_J$*, so the two clauses coincide); and (ii)* realizability*: for every* $g \in G$*, the map* $\psi_g := T^{-1} \circ g \circ T$ *belongs to* $\mathrm{Aut}(\mathcal{Z})_{\Pi,C}$*. Then*

$$V^B(\Pi, C) = \{h \circ T : \ h \in L^\infty(T(\mathcal{Z})), \ h \circ g = h \text{ a.e. on } Y \text{ for every } g \in G\}.$$

Ingredient (i) is an "identifiable up to G" theorem, exactly what the literature produces for each model family. It gives the inclusion $\supseteq$: a G-invariant function takes the same expectation on equivalent instances. Ingredient (ii) gives $\subseteq$: if every g is realized by a symmetry of the class, an identified function must be G-invariant. The proof is in Appendix A. The scalar analogy from the introduction is exactly the gap between (i) and (ii): a theorem saying equivalent solutions differ by $t \mapsto \pm t$ does not prove that only even functions are identified unless some admissible model realizes $t \mapsto -t$.

3 Why normalization is necessary

Theorem 3 (Collapse without normalization)**.** *Let the chart take values in a box* $Y = I_1 \times \cdots \times I_m$ *of nontrivial open intervals, and let* $G \supseteq \prod_k \mathrm{Homeo}_+(I_k)$ *be any group containing all coordinatewise strictly increasing bijections of* Y *(for instance, enlarged by the coordinate permutations when the intervals coincide). Every bounded function on* Y *that is invariant under* G *at every point is constant, and every bounded Borel function that is invariant under each element of* G *almost everywhere is almost everywhere constant. Hence in a model class in which every element of* G *is realized by a symmetry,* $V^B = \mathbb{R} \cdot \mathbf{1}$*.*

The pointwise statement is transitivity: any two points of Y are connected by a coordinatewise increasing bijection. The almost-everywhere statement, the one the definition of V^B requires, uses the countably many rational translations inside G and their ergodicity (Appendix B). This is the algebraic form of the observation that unsupervised disentanglement needs inductive biases (Locatello et al., 2019); Horan et al. (2021) give a complementary route, geometric conditions (local isometry

Table 2: The role of each assumption in Theorem 5.

Assumption	Role	Can it be dropped?
(A1) injective decoder	witnesses exist	no; classical counterexamples
(A2) factorial family, product T	the realizing maps decompose coordinatewise	no; Appendix C.4 counterexample
(A3) regularity, variability	the affine relation between charts	no; it is the engine
(A4) strong identifiability	affine $\Rightarrow$ monomial	derivable on bounded and half-line ranges; fails for Gaussian location on $\mathbb{R}^m$ (Theorem 6)
(A5) monotone C^1 charts	invert T^I, build ψ_g	structural: it makes $\psi_{g,I}$ well defined
(A6) normalization and fixed range Y	kills scale and shift; makes invariance well posed on one domain	normalization: no (Theorem 3), but free (Lemma 4); the range defines the stratum, and the conclusion changes with it

and non-Gaussian factors) under which disentanglement becomes possible. An exact statement must therefore begin by normalizing the statistics; the content of an identifiability theorem lives in what the normalization leaves.

4 The normalized iVAE: exact identified content

4.1 Assumptions, and the role each one plays

The identifiable VAE (Khemakhem et al., 2020) models $p(z \mid s)$ as a conditionally factorial exponential family. We state the assumptions explicitly and record, for each, whether it is a normalization that costs nothing, a structural restriction, or a choice of stratum (Table 2).

(A1) The decoder f is injective.

(A2) The channel is conditionally factorial with scalar statistics: $p(z \mid s) = \prod_{i=1}^{m} Q_i(z_i)\, Z_i(s)^{-1} \exp\big(T_i(z_i)\lambda_i(s)\big)$. Product form of $T = (T_1(z_1), \ldots, T_m(z_m))$ is part of this definition.

(A3) The regularity and variability conditions of Theorem 1 of Khemakhem et al. (2020): the T_i are differentiable almost everywhere, and there are auxiliary values $s_0, \ldots, s_m$ for which the $m \times m$ matrix with columns $\lambda(s_k) - \lambda(s_0)$ is invertible. (Their remaining condition, on the characteristic function of the observation noise, holds trivially here because $X = f(Z)$ is observed without noise.)

(A4) *Strong identifiability*: any two observationally equivalent instances of the class have statistics related by a monomial affine map (one nonzero entry per row and column). This is a hypothesis on the class. Theorems 2 and 3 of Khemakhem et al. (2020) deliver conclusions of this type under conditions on the statistics that differ from ours (their scalar-statistic case assumes non-monotone statistics, which (A5) excludes), so we do not derive (A4) from them; Theorem 6 shows where it is automatic and exhibits a class on $\mathbb{R}^m$ where it fails.

(A5) Each $t_i := T_i^I$ is a C^1 diffeomorphism of an open interval $\mathcal{Z}_i$ onto an open interval Y_i: strictly monotone with a derivative that never vanishes, so that t_i^{-1} is C^1 as well and densities transform by the usual change of variables.

(A6) *Normalization and range*: $\mathbb{E}[T_i] = 0$ and $\mathrm{Var}[T_i] = 1$ under the marginal law of Z for every i, and the range $Y := T^I(\mathcal{Z}) = \prod_i Y_i$ is fixed across the class. The leading case is $Y = \mathbb{R}^m$.

Lemma 4 (Coordinatewise normalization costs nothing)**.** *If $0 < \sigma_i^2 = \mathrm{Var}[T_i] < \infty$, replacing T_i by $(T_i - \mu_i)/\sigma_i$ and λ_i by $\sigma_i\lambda_i$, and absorbing the z-free factor $e^{-\mu_i\lambda_i(s)}$ into the normalizer, leaves the conditional density and the observed law unchanged and preserves (A1) to (A5). Hence the moment part of (A6) can be imposed in every equivalence class of a class closed under affine recodings of the statistics.*

The full whitening $\mathrm{Cov}[T] = I_m$ that one might expect instead is *not* free: it mixes coordinates and leaves the class of coordinatewise charts unless the covariance is already diagonal. Fixing the range Y is also not free; it selects a stratum of the class, and Section 4.4 shows the answer depends on it.

The chart-pinned class. Let $B_m = S_m \ltimes (\mathbb{Z}/2)^m$ be the group of signed permutations and $G_Y := \{g \in B_m : g \cdot Y = Y\}$ the ones preserving the range. Fix a reference instance I satisfying (A1) to (A6). Its class $C = C_I$ consists of *all* instances $J \sim_\Pi I$ that satisfy (A1) to (A6), have the same decoder image $f^J(\mathcal{Z}) = f^I(\mathcal{Z})$, and one of whose admissible charts is a relabeling of the reference chart, $T^J = g_J \circ T^I$ for some $g_J \in G_Y$ (charts of an exponential family are determined only up to such relabelings). Assumption (A4) is imposed on this class as a whole. Nothing else is put into the definition; that the transported instances $\psi_{g,I} \cdot J$ of (1) below belong to C is proved (Lemma 14), not assumed, and that proof is ingredient (ii). What the definition excludes matters as much: the class is *not* closed under arbitrary coordinatewise monotone reparameterizations of Z, because such a map sends any instance to an equivalent one with a new chart, and Section 6 shows that admitting all of them makes each a symmetry and collapses the identified content to constants. Pinning the chart fixes the latent parameterization up to G_Y, which is the sense in which "a function of the latents" can be identified at all; Theorem 10 lifts it to a global statement.

4.2 Where the signed permutations come from

A normalization by first and second moments alone does not give the signed permutations: whitening $T' = AT + c$ gives $c = 0$ and $AA^\top = I$, the orthogonal group $O(m)$, and that ambiguity is realized by a concrete family (Theorem 6(c)); the coordinatewise normalization of (A6) constrains even less on its own. The complete chain has four steps. (1) Weak identifiability (A3) gives, for equivalent instances I, J with witness ψ, an affine relation $T^J \circ \psi = A\,T^I + c$ almost surely. (2) Strong identifiability (A4) forces A to be monomial. (3) The coordinatewise normalization (A6) forces $c = 0$ and the nonzero entries of A to be ± 1, so $A \in B_m$. (4) Fixing the range forces $A \cdot Y = Y$, so $A \in G_Y$. For $Y = \mathbb{R}^m$ every signed permutation preserves the range and $G_Y = B_m$; for a half-line range no sign flip does, and $G_Y \subseteq S_m$.

4.3 The exact equality

For each $g \in G_Y$ define the *global automorphism*

$$\psi_{g,I}(z) := (T^I)^{-1}\big(g \cdot T^I(z)\big). \tag{1}$$

Because T^I has product form, $\psi_{g,I}(z)_i = t_i^{-1}\big(\varepsilon_i\, t_{\pi(i)}(z_{\pi(i)})\big)$ for $g = (\varepsilon, \pi)$: a permutation of coordinates followed by coordinatewise monotone maps. Lemma 14 in Appendix C checks that $\psi_{g,I} \cdot J$ again satisfies (A1) to (A6) and has a chart in $G_Y \cdot T^I$, so $\psi_{g,I}$ maps the class to itself, and it preserves the observed law because relabeling the latents never changes the law of (X, S). Without product form the construction fails: for $T(z_1, z_2) = (z_1 + z_2, z_2)$, injective but not product form, the coordinate swap transports to $z \mapsto (-z_1, z_1 + z_2)$, which does not preserve a factorial class (Appendix C.4).

Theorem 5 (Exact identified content of the normalized iVAE, class-local)**.** *Under (A1) to (A6), for every reference instance I and its chart-pinned class $C = C_I$,*

$$V^B(\Pi, C) = \{h \circ T^I :\ h \text{ bounded Borel on } Y,\ h \circ g = h \text{ for all } g \in G_Y\}.$$

In the leading case $Y = \mathbb{R}^m$, $G_Y = B_m$ and $V^B(\Pi, C)$ is the set of bounded Borel functions of the orbit coordinates $q := \mathrm{sort}(|T_1^I|, \dots, |T_m^I|)$, equivalently of the elementary symmetric polynomials e_k of $((T_1^I)^2, \dots, (T_m^I)^2)$.

The proof applies Theorem 2 with $G = G_Y$: witnesses come from the four-step chain of Section 4.2, realizability from (1). Two remarks on reading the statement. First, the generators e_k are unbounded when $Y = \mathbb{R}^m$ and are not themselves elements of V^B; what is identified is every bounded function of them, such as $\tanh(e_k)$ or $\mathbf{1}\{\sum_i T_i^2 \le r\}$, and an unbounded invariant is identified almost surely as the limit of its truncations, each of which lies in V^B. Second, the statement is a decision procedure: Table 1 is its $m = 2$ instance, and a coordinate i is identified (every bounded function of T_i lies in V^B) exactly when every element of the residual group fixes i with sign $+1$. Under B_m no coordinate is identified for any $m \ge 1$, because a single sign flip suffices; what is identified is every bounded function of the multiset $\{|T_1|, \dots, |T_m|\}$. Range geometry alone can remove the ambiguity: G_Y is trivial exactly when $Y_i \ne -Y_i$ for every i and $Y_j \notin \{Y_i, -Y_i\}$ for $i \ne j$, and then every bounded function of T^I is identified.

4.4 WHAT THE STRONG-IDENTIFIABILITY ASSUMPTION IS DOING

Theorem 6 (Range geometry decides the status of (A4))**.** *Assume (A1) to (A3), (A5), (A6), full support of the latent law, and continuous charts and witnesses.*

(a) ***Bounded range.*** *If every Y_i is a bounded interval, the affine relation between equivalent charts is automatically a signed permutation: (A4) is a consequence, and Theorem 5 holds with (A4) deleted. The realizable group is G_Y; for a symmetric cube $Y = (-b, b)^m$ it is all of B_m.*

(b) ***Half-line range.*** *If every Y_i is a half-line (ℓ_i, ∞) with $\ell_i < 0$ (so that the mean-zero normalization is attainable), the relation is a pure permutation with positive scalings, which the normalization sets to one: no sign flip is possible, and signs are identified by geometry alone.*

(c) ***Full range, Gaussian location.*** *In the conditionally factorial Gaussian location class ($T(z) = z$, $Y = \mathbb{R}^m$, arbitrary conditional variances and any nondegenerate law of the conditional means), every instance admits a continuum of non-monomial class- and normalization-preserving ambiguities, so (A4) cannot be dropped. At the isotropic instance the ambiguity set is exactly $O(m)$ and $V^B(\Pi, C)$ is exactly the bounded functions of $\|T^I\|$.*

The bounded case uses that an invertible affine map carrying a box onto a box permutes edges, hence is monomial; the half-line case uses that a linear bijection of the nonnegative orthant permutes its extreme rays; the Gaussian case sandwiches an orthogonal matrix between positive diagonal scalings that restore the diagonal covariance and the normalization (Appendix D); that rotations are an ambiguity of the Gaussian location family is the content of Proposition 1 of Khemakhem et al. (2020), and part (c) adds that the ambiguity survives every anisotropic configuration and identifies the surviving content exactly at the isotropic one. The design message: a modeler free to choose the statistics can choose bounded or one-sided ones, in which case exact identifiability holds with (A4) derived and, in the one-sided case, without any sign normalization. Choosing bounded statistics specifies a different exponential family, not a re-chart of an unbounded one.

Example 7 (A family satisfying everything at once)**.** *Take $T_i(z_i) = \sqrt{3}\tanh z_i$, so $Y = (-\sqrt{3}, \sqrt{3})^m$, and let the normalized marginal of T_i have density $\rho_{\epsilon_i}(t) = \frac{1}{2\sqrt{3}}\big(1 + \epsilon_i(t^3 - \frac{9}{5}t)\big)$ on $(-\sqrt{3}, \sqrt{3})$ with $0 < \epsilon_i < 5/(6\sqrt{3})$ pairwise distinct: mean 0, variance 1, third moment $\frac{108}{175}\epsilon_i$. These marginals are realized exactly by a factorial channel with independent binary auxiliary signs S_i, base density $\rho_{\epsilon_i}(t)/\cosh(a_i t)$, tilts $\pm a_i$, and $P(S_i = +) = (1 + c_i)/2$ with $c_i := \int \rho_{\epsilon_i} \tanh(a_i t)\, dt$, which makes the marginal of T_i exactly ρ_{ϵ_i}; the natural-parameter differences $2a_i e_i$ give full-rank variability. With an injective decoder, (A1) to (A3), (A5), (A6) hold and (A4) is derived by Theorem 6(a); the cube is B_m-invariant, so $G_Y = B_m$ (Appendix E).*

5 GETTING INDIVIDUAL COORDINATES BACK

Theorem 5 says individual coordinates are not identified under the standard normalization. A slightly stronger normalization changes that, and the mechanism is elementary. After standardization, two equivalent fits can differ by $T_i' = \varepsilon_i T_{\pi(i)}$. Third moments transform as $\mathbb{E}[(T_i')^3] = \varepsilon_i \kappa_{\pi(i)}$, where $\kappa_j := \mathbb{E}[T_j^3]$ is the skewness of coordinate j.

Theorem 8 (Skewness normalization selects a representative)**.** *Under the hypotheses of Theorem 5 with $Y = \mathbb{R}^m$, suppose the skewnesses satisfy $\kappa_j \neq 0$ for all j and $|\kappa_j| \neq |\kappa_k|$ for $j \neq k$. Then the rule "flip signs so that every skewness is positive, then order coordinates by decreasing skewness" is satisfied by exactly one element of each B_m-orbit. Let $C_{\mathrm{skew}} \subset C$ be the subclass of instances whose chart satisfies the rule. Its residual group is trivial and every coordinate is identified on it: $V^B(\Pi, C_{\mathrm{skew}}) = \{h \circ T^I : h \text{ bounded Borel}\}$ for the rule-satisfying reference chart T^I. If instead fixed targets $\mathbb{E}[T_i^3] = c_i$ are imposed, the subclass of instances meeting the targets has residual group $H_c = \{(\varepsilon, \pi) \in B_m : \varepsilon_i c_{\pi(i)} = c_i \ \forall i\}$, the targets are attainable on an orbit with skewnesses κ exactly when $\{|\kappa_j|\} = \{|c_i|\}$ as multisets, and equal positive targets leave $H_c = S_m$.*

This turns "which quantities are identified" into "which normalization do I need": a "No" row of Table 1 becomes "Yes" once the skewness rule is added, as a moment constraint or a post-hoc relabeling. Two further pieces make it usable with finite data.

Theorem 9 (Finite-sample rule)**.** *Let Z have standardized coordinates bounded by $L \geq 1$ with skewnesses κ_j, and define the margin $\delta := \min\{\min_j |\kappa_j|, \frac{1}{2}\min_{j\neq k} ||\kappa_j| - |\kappa_k||\} > 0$. Given n independent samples of $g_0 \cdot Z$ for an unknown $g_0 \in B_m$, the empirical version of the rule (empirical skewness, flip, sort) returns the correct signed permutation except with probability at most $6m\exp(-n\delta^2/(3600L^{12}))$. If the samples are perturbed by a bounded representation error $|R_j| \leq \eta \leq \delta/(252(L+1)^5)$, the same holds with δ replaced by $\delta/2$.*

The constant is unoptimized and conservative ($n\delta^2 \gtrsim 30$ sufficed in our designs), but the shape is right: read backwards, it prescribes statistics whose skewness margin satisfies $n\delta^2 \gtrsim L^{12}\log m$ before training. It also explains our negative result below: the coordinate the rule failed to stabilize had $|\kappa| = 0.24$ next to a neighbor at 0.54, a small margin relative to the representation error of poorly converged restarts.

A certificate that abstains. For a fitted representation with a known bounded range, evaluated on n held-out points, compute empirical-Bernstein confidence intervals (Maurer & Pontil, 2009) for the first three raw moments of each coordinate at level $\alpha/(3m)$, propagate them by outward-rounded interval arithmetic to the third central moment (the sign decision) and to the cross-multiplied ordering inequalities (the order decisions, no division), and certify a decision only if its interval excludes zero. With probability at least $1-\alpha$, every certified decision equals the population rule applied to that representation (Appendix F.1); otherwise the procedure abstains. It certifies normalization decisions, not closeness to the truth.

6 The global picture

Partition the class into equivalence classes, $\mathcal{M} = \bigcup_\alpha C_\alpha$. Then $V^B(\Pi, \mathcal{M}) = \big(\bigcap_\alpha V^A(\Pi, C_\alpha)\big) \cap \mathrm{Inv}(\mathrm{Aut}(\mathcal{Z})_{\Pi,\mathcal{M}})$ and $\bigcap_\alpha V^B(\Pi, C_\alpha) \subseteq V^B(\Pi, \mathcal{M})$, because global symmetries are symmetries of every class (Appendix G). Whether the reverse inclusion holds depends on how the class is closed; both extremes are theorems.

Theorem 10 (Rigid closure: global exactness)**.** *Fix a product-form chart T satisfying (A5) and (A6) with $Y = \mathbb{R}^m$, and let $\mathcal{M}_T$ be the instances whose chart is $g_I \cdot T$ for some $g_I \in B_m$, closed under B_m-relabeling of the chart but not under monotone reparameterization of Z. Under (A1) to (A4), $V^B(\Pi, \mathcal{M}_T) = \{h \circ T : h$ bounded and B_m-invariant$\}$.*

Theorem 11 (Monotone closure: global collapse)**.** *Let $\mathcal{M}$ contain a reference instance I with product-form bijective chart $T^I : \mathcal{Z} \to \mathbb{R}^m$ such that $P_Z^I \circ (T^I)^{-1}$ is equivalent to Lebesgue measure, and let $\mathcal{M}$ be closed under coordinatewise strictly monotone reparameterizations of Z. Then $V^B(\Pi, \mathcal{M}) = \mathbb{R}\cdot\mathbf{1}$, and also $V^B(\Pi, C) = \mathbb{R}\cdot\mathbf{1}$ for the class C of I inside $\mathcal{M}$.*

The collapse proof conjugates every product of increasing homeomorphisms through T^I, reduces to rational translations, and uses their ergodicity coordinate by coordinate (Appendix G). Read together, the two theorems say what the model class must look like for an exact statement to exist: pinning the chart preserves the content, class by class (Theorem 5) and globally (Theorem 10), while admitting all coordinatewise monotone reparameterizations makes each of them a symmetry of every class and destroys the content, class-locally as well as globally (Theorem 11). The choice is not a technicality: it is the difference between an identifiability theory of latent functions and no theory at all.

Other families through the same interface. The test of Theorem 2 mentions neither exponential families nor the iVAE: each identifiability theorem in the literature supplies ingredient (i) for its own class, and ingredient (ii) must be checked family by family. Appendix H works through content–style separation (von Kügelgen et al., 2021) and CITRIS (Lippe et al., 2022) at the level of the interface, and notes why contrastive learning on the hypersphere (Zimmermann et al., 2021) identifies no single-sample function of the latents (its witnesses form $O(m)$), while score- and flow-based families (Roeder et al., 2021; Hälvä et al., 2021) are outside what we have verified.

7 Experiments with fitted models

The algebra of Sections 4 and 5 is exact; Figure 1 illustrates it once. The questions that matter for practice are different: does a fitted iVAE land on the B_m-orbit, are identified functions the stable

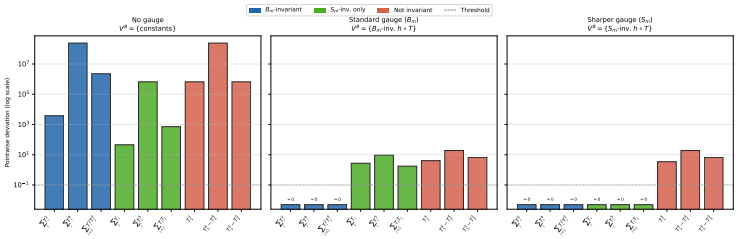


Figure 1: Statistic-level illustration of the algebra ($m = 3$, standardized statistics with distinct skewness). For nine test functions, the largest deviation $\max_g \mathbb{E}|h(g \cdot T) - h(T)|$ over the residual group. Left: no normalization, every non-constant function varies (Theorem 3). Center: standard normalization, only the B_m-invariant functions (blue) are stable (Theorem 5). Right: residual group reduced to S_m, as under a sign normalization, and the S_m-invariant functions (green) also become stable (Theorem 8). The panel titles use "gauge" for normalization.

Table 3: Fitted iVAE, $m = 2$, $n = 8000$, eight restarts, held-out evaluation. Top: cross-restart standard deviation of each functional divided by its across-observation standard deviation, the mean over restarts, and the value on the true latents; for the unbounded $\min(|\hat{z}_1|, |\hat{z}_2|)$, "identified" means through its bounded transforms. Bottom: $\tanh(\hat{z}_i)$ before and after the data-driven skewness rule, and, for coordinate 1, the number of restarts agreeing with the modal (fitted coordinate, sign) assignment before and after the rule.

Functional	Class	normalized SD	restart mean	true value
$\tanh(\hat{z}_1^2 + \hat{z}_2^2)$	identified	0.051	0.672	0.656
$\min(\lvert\hat{z}_1\rvert, \lvert\hat{z}_2\rvert)$	identified	0.067	0.494	0.520
$\max(\hat{z}_1, \hat{z}_2)$	not (S_m-only)	0.181	0.562	0.319
$\tanh(\hat{z}_1)$	not	0.129	0.043	-0.085
$\tanh(\hat{z}_1 + \hat{z}_2)$	not	0.150	0.007	-0.126
		raw SD	skew-normalized SD	modal agreement
coordinate 1, $\tanh(\hat{z}_1)$		0.078	0.011	$4/8 \to 8/8$
coordinate 2, $\tanh(\hat{z}_2)$		0.066	0.033	

ones across retraining, and does the skewness rule work on fitted latents? We answer them on a small pipeline that runs on one CPU: $m = 2$, a conditionally factorial Gaussian location-scale prior over six auxiliary values (variability rank $4/4$, smallest singular value 0.21, marginal skewnesses $(1.06, 0.50)$), a random injective three-layer decoder to ten dimensions with noise 0.05, a standard iVAE with diagonal Gaussian encoder trained by ELBO, three master datasets evaluated through nested prefixes of size $n \in \{50, 125, 500, 2000, 8000\}$, eight optimizer seeds per cell (120 fits), and all metrics on a fixed held-out set. The Gaussian family has two statistics per coordinate, outside the literal scope of Theorem 5, but its residual group after standardization is B_m by the Gaussian analysis of Sorrenson et al. (2020); a scalar-statistic family is fitted at the end of the section. Full protocol and the $m = 3$ run are in Appendix I.

The symmetry is real in fitted models. Aligning the fitted latents to the truth by the best element of B_2 reduces the per-coordinate error from 2.0 to 2.3 (unaligned, at every n) to between 0.03 and 0.14 for the best restart of each of the fifteen cells (correlation with the truth 0.93 to 0.98), and the best-aligning group element is *not* the identity in 108 of 120 fits. Labels and signs scramble across restarts, consistent with the ambiguity that Theorem 5 predicts (fitted models are only approximately equivalent, so this is evidence, not a verification). One pre-stated prediction failed and we report it: the *mean* aligned error did not decrease with n (cell means between 0.19 and 0.31 with no trend; per-n means of the best restart 0.06, 0.04, 0.04, 0.10, 0.05, several times the exact-posterior floor of 0.0078), and the validation ELBO did not pick the recovering restarts (pooled rank correlation with aligned error -0.26, sign inconsistent across cells). The bottleneck at these scales is the amortized variational fit, not the sample size or the identification theory; the next two findings survive that regime.

Identified functions are the stable interface. Table 3 (top) evaluates five functionals of the standardized fitted latents across the eight restarts of one cell ($n = 8000$), without using ground truth for anything but the last column; it is an illustration on one cell, not a survey. The two identified functionals are roughly two to three times more stable across restarts after scale normalization (ratios 1.9 to 3.5) and their means sit near the true values; the three non-identified ones are unstable and far from the truth; whoever consumes a "No" row depends on which restart happened to run.

The skewness rule restores coordinates. Applying the rule of Theorem 8 to the fitted latents, using only their empirical skewness, reduces the cross-restart standard deviation of the strongly skewed coordinate seven-fold and makes all eight restarts agree on which fitted coordinate is coordinate one and with which sign, up from four of eight (Table 3, bottom). The weakly skewed coordinate improves by a factor two; in the family below with skewnesses $(-0.24, -0.54)$ it is not stabilized in poorly converged restarts, the margin effect that Theorem 9 quantifies. The certificate of Section 5 was run on 200 exact-up-to-B_m representations with n from 250 to 256,000 and on the 16 fitted models of the scalar-statistic family: it never certified a wrong decision in the 216 runs, it abstained on all 16 fitted models (whose orientations, measured against the truth, disagreed in 16 of 16 cases), and on exact representations it abstained through $n = 16{,}000$ and certified every decision at $n = 64{,}000$ and 256,000.

A scalar-statistic family. Because the Gaussian pipeline has two statistics per coordinate, we also fit a family with the channel structure of (A2), $p(z_i \mid u) \propto \exp(\lambda_{u,i} \tanh z_i)\varphi(z_i)$, with the matching prior and learnable natural parameters ($k = 1$, six auxiliary values, variability rank $2/2$, distinct nonzero skewness). It is a sanity check of optimization under the theorem's channel structure, not a literal test of Theorem 5, since the fitting does not enforce the pinned range of (A6). Over sixteen fits the best-aligning element is not the identity in sixteen of sixteen, the best restart reaches correlation 0.98 with the truth after alignment, and the skewness rule partially stabilizes the strongly skewed coordinate (pair consistency from 64% to 79%) and not the weakly skewed one.

8 Related work, limitations, and conclusion

The iVAE theorem (Khemakhem et al., 2020) and its relatives in nonlinear ICA (Hyvärinen & Morioka, 2016; 2017; Hyvärinen et al., 2019) give "up to symmetry" statements; the $k = 2$ Gaussian within-block classification is due to Sorrenson et al. (2020). Content–style separation (von Kügelgen et al., 2021), contrastive inversion (Zimmermann et al., 2021), temporal methods (Yao et al., 2022), interventional methods with polynomial decoders (Ahuja et al., 2023), CITRIS (Lippe et al., 2022), and mechanism sparsity (Lachapelle et al., 2022) each supply ingredient (i) for their own class; function-class approaches (Gresele et al., 2021; Buchholz et al., 2022) restrict the mixing instead. Impossibility results (Locatello et al., 2019) and the geometric possibility conditions of Horan et al. (2021) bracket the setting of Theorem 3.

Limitations. The exact results are for the chart-pinned normalized iVAE class with scalar statistics under (A1) to (A6); (A4) is a genuine hypothesis on the full range, where the Gaussian-location class shows it cannot be dropped in general. Vector statistics ($k \geq 2$), non-exponential-family channels, score- and flow-based models, and CITRIS are handled only at the interface level. The finite-sample rule has a conservative constant, and the certificate certifies normalization decisions, not distance to the truth. The fitted-model study is small; its clearest finding is negative: at these scales the amortized ELBO neither recovers the orbit reliably nor selects the restarts that do.

Conclusion. An identifiability theorem is an upper bound on ambiguity; turning it into a statement about what the data determine requires realizing every symmetry inside the class. For the chart-pinned, scalar-statistic normalized iVAE class the identified content is exactly the bounded functions invariant under the range-preserving signed permutations. The two-ingredient test tells other families what to verify; the collapse theorem says why normalization is not optional; the skewness rule buys back coordinates; the fitted-model experiments show which quantities survive retraining.

Reproducibility statement

Complete proofs of every stated result are in Appendices A to G, with the assumptions listed in Section 4. Appendix I specifies the data-generating processes, architectures, training schedules, evaluation sets, and metrics of every experiment; random seeds are fixed inside the anonymized supplementary code, which reproduces every reported table and figure (the 120-fit grid, the scalar-statistic family, and the certificate runs) on a single CPU in minutes.

Ethics statement

The work is theoretical and uses only synthetic data. It aims to make downstream uses of learned representations better defined; we see no direct negative societal consequences.

AI use statement

Large language models were used substantially in preparing this manuscript: drafting and editing text, preparing LaTeX, searching and checking the literature, writing and reviewing experimental code, and cross-checking derivations, proofs, and numerical results. The authors posed the research questions, chose the statistical models and assumptions, decided every claim, derived and verified the mathematical statements and proofs and the experimental results, and take full responsibility for the content. Every bibliography entry was checked against its primary record.

References

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 372–407. PMLR, 2023.

Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function classes for identifiable nonlinear independent component analysis. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Morris L. Eaton. *Group Invariance Applications in Statistics*, volume 1 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, 1989.

Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvärinen. Disentangling identifiable features from noisy data with structured nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Daniella Horan, Eitan Richardson, and Yair Weiss. When is unsupervised disentanglement possible? In *Advances in Neural Information Processing Systems*, volume 34, pp. 5150–5161, 2021.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 460–469. PMLR, 2017.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E. Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Proceedings of the 1st Conference on Causal Learning and Reasoning*, pp. 428–484. PMLR, 2022.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, 3rd edition, 2005.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 13557–13603. PMLR, 2022.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4114–4124. PMLR, 2019.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Geoffrey Roeder, Luke Metz, and Diederik P. Kingma. On linear identifiability of learned representations. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9030–9039. PMLR, 2021.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN). In *International Conference on Learning Representations*, 2020.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12979–12990. PMLR, 2021.

A Proof of Theorem 2

Proof. ($\supseteq$) Let $h \circ g = h$ a.e. on Y for every $g \in G$ and $\varphi := h \circ T$. For $J \sim_\Pi J'$ in C, ingredient (i) gives ψ with $P_Z^{J'} = \psi_\# P_Z^J$ and $\varphi \circ \psi = h(g_\psi \circ T) = h \circ T = \varphi$ a.s. (the exceptional null set of $h \circ g_\psi = h$ pulls back to a null set of $\mathcal{Z}$ because T carries the latent law to the reference measure on Y up to equivalence), so $\int \varphi \, dP_Z^{J'} = \int \varphi \circ \psi \, dP_Z^J = \int \varphi \, dP_Z^J$ and $\varphi \in V^A(\Pi, C)$. For any $\psi_0 \in \mathrm{Aut}(\mathcal{Z})_{\Pi,C}$ and $J \in C$, the instance $\psi_0 \cdot J$ lies in C and ψ_0 is a relabeling realizing the pair $(J, \psi_0 \cdot J)$, so the second clause of (i) gives $\varphi \circ \psi_0 = h \circ g_{\psi_0} \circ T = \varphi$ a.s. Hence $\varphi \in V^B(\Pi, C)$. ($\subseteq$) Let $\varphi \in V^B(\Pi, C)$ and write $\varphi = h \circ T$ with $h := \varphi \circ T^{-1}$ on $T(\mathcal{Z})$, bounded and measurable because T is a measurable injection with measurable inverse. For $g \in G$, ingredient (ii) gives $\psi_g \in \mathrm{Aut}(\mathcal{Z})_{\Pi,C}$, so $\varphi \circ \psi_g = \varphi$ a.s., that is $h(g \cdot T(z)) = h(T(z))$ for a.e. z; pushing forward by T gives $h \circ g = h$ a.e. on Y, for each g separately, which is the stated form. □

B Proof of Theorem 3

Proof. For $y, y' \in Y$ and each coordinate k choose $g_k \in \mathrm{Homeo}_+(I_k)$ with $g_k(y_k) = y'_k$: a translation if $I_k = \mathbb{R}$, the piecewise linear increasing bijection fixing the endpoints and sending y_k to y'_k if I_k is bounded, and the analogous construction on half-lines. Then $g = (g_1, \ldots, g_m)$ satisfies $g \cdot y = y'$, so a pointwise G-invariant h has $h(y) = h(y')$: h is constant. For the almost-everywhere statement, fix increasing homeomorphisms $\phi_k : I_k \to \mathbb{R}$ and, for each coordinate k and rational r, the element $g_{k,r} \in \prod_j \mathrm{Homeo}_+(I_j)$ that acts on coordinate k by $\phi_k^{-1}(\phi_k(\cdot) + r)$ and fixes the others. If $h \circ g = h$ a.e. for every $g \in G$, then in the coordinates $w = (\phi_1(y_1), \ldots, \phi_m(y_m))$ the function $\tilde{h}(w) := h(y)$ is invariant a.e. under every rational translation of every coordinate; the countably many exceptional null sets have a null union, so by Fubini almost every one-dimensional slice of $\tilde{h}$ is a.e. invariant under rational translations of $\mathbb{R}$, hence a.e. constant because the rational translations act ergodically on $(\mathbb{R}, \lambda)$; iterating over coordinates, $\tilde{h}$ and hence h are a.e. constant. If every element of G is realized by a symmetry of the class, every $\varphi \in V^B$ is a.e. invariant under G by definition, hence a.e. constant. □

C Proof of Theorem 5

C.1 Witnesses

Lemma 12 (Injective strong equivalence)**.** *If $I \sim_\Pi I'$, both decoders are injective with the same image $f(\mathcal{Z}) = f'(\mathcal{Z})$, and both latent laws have positive densities on $\mathcal{Z}$, there is a measurable bijection $\psi : \mathcal{Z} \to \mathcal{Z}$ with $P_Z^{I'} = \psi_\# P_Z^I$, $\kappa' = \kappa \circ \psi^{-1}$ a.s. and $f' = f \circ \psi^{-1}$.*

Proof. Observational equivalence gives $P_X^I = P_X^{I'}$. Define $\psi := f'^{-1} \circ f$ on $\mathcal{Z}$; f'^{-1} is measurable on $f'(\mathcal{Z}) = f(\mathcal{Z})$ by the Lusin–Souslin theorem, so ψ is a measurable bijection of $\mathcal{Z}$. For measurable B, $P_Z^{I'}(B) = P_X^{I'}(f'(B)) = P_X^I(f'(B)) = P_Z^I(\psi^{-1}(B))$. Matching the conditional laws of S given X on the common support gives the channel transport, and $f' \circ \psi = f$ by construction. □

Lemma 13 (Witnesses lie in G_Y)**.** *Under (A1) to (A6), for every $J, J' \in C$ there is a measurable bijection ψ with $J' = \psi \cdot J$, and every measurable bijection ψ with $J' = \psi \cdot J$ satisfies $T^I \circ \psi = g \cdot T^I$ a.s. for some $g \in G_Y$.*

Proof. Step 1 (existence and uniqueness). Lemma 12 gives $\psi = f_{J'}^{-1} \circ f_J$ with $J' = \psi \cdot J$; any ψ' with $J' = \psi' \cdot J$ satisfies $f_{J'} = f_J \circ \psi'^{-1}$, so $\psi' = f_{J'}^{-1} \circ f_J = \psi$ on the support of P_Z^J by injectivity. *Step 2 (affine relation).* Both instances have the same law of (X, S), and by (A3) Theorem 1 of Khemakhem et al. (2020) gives $T^{J'} \circ f_{J'}^{-1} = A\,(T^J \circ f_J^{-1}) + c$ on the observed support for an invertible A; composing with f_J, $T^{J'} \circ \psi = A\,T^J + c$ a.s. *Step 3 (monomial, then signed permutation).* By (A4) A is monomial: $A = P_\pi D$ with D diagonal and nonzero. By (A6) and $\psi_\# P_Z^J = P_Z^{J'}$, each coordinate of $T^{J'} \circ \psi$ has mean 0 and variance 1 under P_Z^J, as does each coordinate of T^J; the right side has mean c_j and variance D_{ii}^2 for $j = \pi(i)$, whence $c = 0$ and $D_{ii} = \pm 1$, so $A \in B_m$. Since both charts have range Y and ψ is a bijection, $A \cdot Y = Y$, so $A \in G_Y$.

Step 4 (pinning). In the chart-pinned class, $T^J = g_J \circ T^I$ and $T^{J'} = g_{J'} \circ T^I$ with $g_J, g_{J'} \in G_Y$; substituting, $g_{J'} \cdot (T^I \circ \psi) = A\, g_J \cdot T^I$, so $T^I \circ \psi = (g_{J'}^{-1} A g_J) \cdot T^I$ with $g_{J'}^{-1} A g_J \in G_Y$. □

C.2 Realizability

Lemma 14 (Product-form global automorphism). *Let $T^I(z) = (t_1(z_1), \dots, t_m(z_m))$ with each t_i a C^1 strictly monotone bijection $\mathcal{Z}_i \to Y_i$, and let $g = (\varepsilon, \pi) \in G_Y$. Then (a) $\psi_{g,I}(z)_i = t_i^{-1}(\varepsilon_i t_{\pi(i)}(z_{\pi(i)}))$, a coordinate permutation composed with coordinatewise strictly monotone maps, well defined because $g \cdot Y = Y$; (b) for every $J \in \mathcal{M}$ with factorized chart $T^J(z) = (s_1(z_1), \dots, s_m(z_m))$, the transported chart $T^J \circ \psi_{g,I}^{-1}$ is factorized up to the coordinate permutation π^{-1}; (c) $\psi_{g,I} \in \mathrm{Aut}(\mathcal{Z})_{\Pi,C}$.*

Proof. (a) $(T^I)^{-1}(y) = (t_1^{-1}(y_1), \dots, t_m^{-1}(y_m))$ and $[g \cdot T^I(z)]_i = \varepsilon_i t_{\pi(i)}(z_{\pi(i)}) \in Y_i$ because $g \cdot Y = Y$; compose. (b) $\psi_{g,I}^{-1} = (T^I)^{-1} \circ g^{-1} \circ T^I$ has the same form with $g^{-1} = (\varepsilon', \pi^{-1})$, $\varepsilon'_j = \varepsilon_{\pi^{-1}(j)}$, so $[\psi_{g,I}^{-1}(w)]_j$ depends on $w_{\pi^{-1}(j)}$ alone and $\tilde{T}_k(w) = s_k([\psi_{g,I}^{-1}(w)]_k)$ is a function of $w_{\pi^{-1}(k)}$ only. (c) Relabeling never changes the law of (X, S): with $Z' = \psi_{g,I}(Z)$, the decoder $f \circ \psi_{g,I}^{-1}$ returns the same X and the channel $\kappa \circ \psi_{g,I}^{-1}$ the same S. The transported instance satisfies (A1) to (A6): the decoder stays injective; the conditional density of Z' given s is the product $\prod_i Q_i(\psi^{-1}(z')_i)\, |\mathrm{d}\psi^{-1}(z')_i / \mathrm{d}z'_{\pi^{-1}(i)}|\, Z_i(s)^{-1} \exp(T_i(\psi^{-1}(z')_i)\lambda_i(s))$ (coordinate i of $\psi^{-1}(z')$ depends on $z'_{\pi^{-1}(i)}$ alone), a conditionally factorial exponential family with statistics $T^J \circ \psi_{g,I}^{-1} = (g_J g^{-1}) \cdot T^I$ (by (b)) and correspondingly relabeled natural parameters, so (A2), (A3) and (A5) hold up to relabeling the indices of the statistics; the moments and the range of the statistics are those of T^J under P_Z^J, so (A6) holds; the decoder image is unchanged; and (A4) is a hypothesis on the whole class, so it requires no check. Its chart is $(g_J g^{-1}) \cdot T^I$ with $g_J g^{-1} \in G_Y$, so $\psi_{g,I} \cdot J$ belongs to the chart-pinned class C; this is the closure of C under the transports, that is, ingredient (ii). Hence $\psi_{g,I}$ maps C onto itself, preserves every observed law, and lies in $\mathrm{Aut}(\mathcal{Z})_{\Pi,C}$. □

Proof of Theorem 5. Apply Theorem 2 with $T = T^I$, $G = G_Y$: ingredient (i) is Lemma 13, ingredient (ii) is Lemma 14(c). For $Y = \mathbb{R}^m$, every signed permutation preserves $\mathbb{R}^m$, so $G_Y = B_m$; the B_m-invariant Borel functions on $\mathbb{R}^m$ are exactly the Borel functions of the orbit map $q = \mathrm{sort}(|y_1|, \dots, |y_m|)$, which separates orbits, and $\sigma(q) = \sigma(e_1, \dots, e_m)$ because the elementary symmetric polynomials of $(y_1^2, \dots, y_m^2)$ determine the multiset $\{y_i^2\}$ and hence q. □

C.3 Proof of Lemma 4

The conditional density $Q_i(z_i) Z_i(s)^{-1} \exp(T_i(z_i)\lambda_i(s))$ is unchanged under $T_i \mapsto (T_i - \mu_i)/\sigma_i$, $\lambda_i \mapsto \sigma_i \lambda_i$, once $e^{-\mu_i \lambda_i(s)}$ is absorbed into $Z_i(s)$, so the observed law is unchanged. Injectivity of f (A1), factoriality and product form (A2), differentiability and the rank of the natural-parameter differences (A3, the differences scale by σ_i per row), monomiality of witnesses (A4, conjugation by a diagonal scaling preserves the monomial pattern), and monotone bijectivity of the charts (A5) are all preserved. □

C.4 Counterexample: injectivity without product form

Take $m = 2$, $T(z_1, z_2) = (z_1 + z_2, z_2)$, injective and not product form, and g the coordinate swap. Then $\psi_g(z) = T^{-1}(z_2, z_1 + z_2) = (-z_1, z_1 + z_2)$. Its second output depends on both coordinates, and the transported latents are conditionally dependent: for any instance with conditionally independent coordinates of positive conditional variances, $\mathrm{Cov}(\psi_g(Z)_1, \psi_g(Z)_2 \mid s) = -\mathrm{Var}(Z_1 \mid s) \neq 0$, so the transported instance is not conditionally factorial. Realizability fails while (A1), (A3), (A5), (A6) hold.

D Proof of Theorem 6

Throughout, the standing regularity (full support, continuous charts and witnesses) upgrades the a.s. affine relation $T^J \circ \psi = A\,T^I + c$ of (A3) to an identity on all of $\mathcal{Z}$, and ψ is a bijection with $\psi_{\#}P_Z^I = P_Z^J$.

Lemma 15 (Affine bijections between boxes are monomial)**.** *If $\varphi(x) = Ax + c$ with A invertible maps a nondegenerate closed box $B = \prod_i [a_i, b_i]$ onto a nondegenerate closed box B', then A is monomial.*

Proof. Invertible affine maps preserve the face lattice, mapping r-faces to r-faces. The m edges at the vertex $(a_1, \dots, a_m)$ have directions $e_1, \dots, e_m$; their images are the m edges of B' at $\varphi(v)$, whose directions are distinct coordinate vectors. Hence $Ae_i = \alpha_i e_{\pi(i)}$ with $\alpha_i \neq 0$ and π a bijection. For open boxes apply this to closures. □

Proof of (a). By (A6) both charts have range $Y = \prod_i (a_i, b_i)$ and ψ is a bijection, so $A \cdot Y + c = Y$; by Lemma 15, A is monomial. Then, as in Lemma 13 Step 2, the normalization forces $c = 0$ and entries ± 1, so $A \in B_m$ and $A \cdot Y = Y$, so $A \in G_Y$. This is exactly what (A4) was invoked for, so Theorem 5 follows with (A4) deleted; realizability holds for G_Y by Lemma 14. For a symmetric cube every signed permutation preserves Y. □

Proof of (b). With $Y_i = (\ell_i, \infty)$, the recession cone of $\bar{Y}$ is the nonnegative orthant, and $A \cdot Y + c = Y$ implies $A\,\mathbb{R}^m_{\geq 0} = \mathbb{R}^m_{\geq 0}$. A linear bijection of the orthant permutes its extreme rays, so $Ae_i = \alpha_i e_{\pi(i)}$ with $\alpha_i > 0$. Two fixed per-coordinate moments then give $\alpha_i = 1$ and $c = 0$, and the range identity forces $\ell_{\pi(i)} = \ell_i$. No sign flip occurs. □

Proof of (c). Let $p(z \mid s) = \prod_i \mathcal{N}(z_i; \mu_i(s), v_i)$ with $\mathbb{E}\mu(S) = 0$, $\Sigma_\mu := \mathrm{Cov}(\mu(S)) \succ 0$, $M := \mathrm{Cov}(Z) = D_v + \Sigma_\mu$, and normalization $\mathrm{diag}(M) = I$. For $B \in O(m)$ put $\widetilde{M} := D_v^{-1/2} M D_v^{-1/2}$, $D_B := \mathrm{diag}\big((B\widetilde{M}B^\top)_{ii}\big)^{-1}$ and $A_B := D_B^{1/2} B D_v^{-1/2}$. The instance with latents $A_B Z$, means $A_B\mu(s)$, conditional variances $(D_B)_{ii}$ and decoder $f \circ A_B^{-1}$ has conditional covariance $A_B D_v A_B^\top = D_B$ (diagonal, so conditionally factorial with $T_i(z_i) = z_i$), marginal covariance with unit diagonal by the definition of D_B, zero mean, the same observed law, and natural parameters $\lambda^J(s) = D_B^{-1} A_B \mu(s) = D_B^{-1} A_B D_v \lambda^I(s)$, an invertible image of the original ones, so (A3) holds. A positive diagonal sandwich preserves the zero pattern, so A_B is monomial iff B is; non-monomial $B \in O(m)$ exist for $m \geq 2$. Hence a continuum of non-monomial ambiguities survives. At the isotropic instance ($v = v_0\mathbf{1}$, $\Sigma_\mu = \sigma^2 I$, $M = I$) the construction gives $A_B = B$, so every $B \in O(m)$ is realized; conversely any ambiguity A has $c = 0$ by the mean normalization, $\mathrm{Cov}(AZ \mid s) = v_0 AA^\top$ diagonal, and $\mathrm{diag}(AMA^\top) = \mathrm{diag}(AA^\top) = I$, so $AA^\top = I$. Theorem 2 with $G = O(m)$ (witnesses in $O(m)$ by the converse, realizability by $\psi_B(z) = Bz$) gives $V^B(\Pi, C) =$ bounded $O(m)$-invariant functions of T^I, that is bounded functions of $\|T^I\|$. □

E The family of Example 7

Let $\rho_\epsilon(t) = \frac{1}{2\sqrt{3}}(1 + \epsilon(t^3 - \frac{9}{5}t))$ on $(-\sqrt{3}, \sqrt{3})$. Since $|t^3 - \frac{9}{5}t| \leq \frac{6\sqrt{3}}{5}$ on that interval, $\rho_\epsilon > 0$ for $0 < \epsilon < 5/(6\sqrt{3})$; $\int \rho_\epsilon = 1$ and $\int t^2 \rho_\epsilon = 1$ because the perturbation is odd and $\int_{-\sqrt{3}}^{\sqrt{3}} t^2\,dt/(2\sqrt{3}) = 1$; the mean vanishes because of the coefficient $\frac{9}{5}$: $\int t\rho_\epsilon = \frac{\epsilon}{2\sqrt{3}} \int_{-\sqrt{3}}^{\sqrt{3}} (t^4 - \frac{9}{5}t^2)\,dt = \frac{\epsilon}{2\sqrt{3}}\big(\frac{18\sqrt{3}}{5} - \frac{9}{5} \cdot 2\sqrt{3}\big) = 0$; and $\int t^3 \rho_\epsilon = \frac{\epsilon}{2\sqrt{3}} \int_{-\sqrt{3}}^{\sqrt{3}} (t^6 - \frac{9}{5}t^4)\,dt = \frac{\epsilon}{2\sqrt{3}}\big(\frac{2 \cdot 27\sqrt{3}}{7} - \frac{9}{5} \cdot \frac{2 \cdot 9\sqrt{3}}{5}\big) = \epsilon\big(\frac{27}{7} - \frac{81}{25}\big) = \frac{108}{175}\epsilon$. With $T_i(z_i) = \sqrt{3}\tanh z_i$, the chart is a C^1 increasing bijection $\mathbb{R} \to (-\sqrt{3}, \sqrt{3})$. Channel: independent binary auxiliary signs $S = (S_1, \dots, S_m)$, one per coordinate, with $\lambda_i(s) = a_i s_i$ and base density $Q_i(t) = \rho_{\epsilon_i}(t)/\cosh(a_i t)$ in the t coordinate, so $p(t_i \mid s_i = \pm) = \rho_{\epsilon_i}(t) e^{\pm a_i t}/(\cosh(a_i t) Z_i(\pm)) = \rho_{\epsilon_i}(t)(1 \pm \tanh(a_i t))/(1 \pm c_i)$ with $c_i := \int \rho_{\epsilon_i} \tanh(a_i t)\,dt \in (-1, 1)$. Taking $P(S_i = \pm) = (1 \pm c_i)/2$, the marginal is $\frac{1}{2}\rho_{\epsilon_i}(1 + \tanh) + \frac{1}{2}\rho_{\epsilon_i}(1 - \tanh) = \rho_{\epsilon_i}$ exactly, and the conditional law of T given $S = s$ is a product, so the channel is conditionally factorial with scalar statistics. With every $a_k \neq 0$, the

natural-parameter differences between s and the vector with the kth sign flipped are $2a_k e_k$, so the m single-flip differences form an invertible matrix and (A3) holds. Distinct ϵ_i give distinct nonzero skewnesses, so Theorem 8 also applies. The chart-pinned class of this instance is saturated by Lemma 14(c).

F Proofs for Section 5

Proof of Theorem 8. Signed permutations preserve means and unit variances, and $\mathbb{E}[(\varepsilon_i T_{\pi(i)})^3] = \varepsilon_i \kappa_{\pi(i)}$. On an orbit $\{g \cdot T\}$, the rule "all skewnesses positive" fixes $\varepsilon_i = \mathrm{sign}(\kappa_{\pi(i)})$, possible since no κ_j vanishes, and "decreasing skewness" fixes π uniquely since the $|\kappa_j|$ are pairwise distinct. So exactly one g per orbit satisfies the rule, the class of rule-satisfying instances has trivial residual group, and Theorem 2 with $G = \{\mathrm{id}\}$ (witnesses trivial by Lemma 13 restricted to rule-satisfying instances, realizability trivial) gives all bounded Borel functions of T^I. For fixed targets, $\mathbb{E}[(\varepsilon_i T_{\pi(i)})^3] = c_i$ for all i iff $\varepsilon_i c_{\pi(i)} = c_i$, which defines H_c; attainability on an orbit requires $\{|\kappa_j|\} = \{|c_i|\}$; equal positive targets give $\varepsilon \equiv 1$ and arbitrary π. □

Proof of Theorem 9. Write μ_j for the skewnesses and observe that coordinate j of $g_0 \cdot Z$ has skewness $\varepsilon_j \mu_{\pi(j)}$. The empirical skewness of a coordinate is its third empirical central moment divided by the $3/2$ power of its empirical variance, with $\mathrm{sign}(0) := +1$; if an empirical variance vanishes the rule's output is arbitrary and counted as a failure, and for $m = 1$ the second minimum in δ is $+\infty$. ***Step 1 (margin).*** If every empirical skewness $\hat{m}_j$ satisfies $|\hat{m}_j - \varepsilon_j \mu_{\pi(j)}| \le e < \delta$, then the empirical sign is correct because $|\mu_{\pi(j)}| \ge \delta > e$, and after flipping, the $|\hat{m}_j|$ are within e of the $|\mu_{\pi(j)}|$, which differ pairwise by at least $2\delta > 2e$, so the sort recovers the canonical assignment. ***Step 2 (Lipschitz).*** With raw moments $A_r = \frac{1}{n}\sum_i Y_i^r$ and population values $(0, 1, \mathrm{skew})$, the empirical skewness is $F(A_1, A_2, A_3) = (A_3 - 3A_1A_2 + 2A_1^3)/(A_2 - A_1^2)^{3/2}$. On $\{|A_1| \le \frac{1}{4}, |A_2 - 1| \le \frac{1}{4}\}$ with $|A_3| \le L^3$ the partial derivatives are bounded by 1.76, $9L^3$, $12L^3$, so F is $42L^3$-Lipschitz in $\max_r |A_r - a_r|$ there. ***Step 3 (concentration).*** Hoeffding for $Y^r \in [-L^r, L^r]$ gives $P(|A_r - a_r| > t) \le 2\exp(-nt^2/(2L^6))$ for $r \le 3$. Take $t = \delta/(42L^3) \le 1/42$; a union bound over 3 moments and m coordinates gives $P(\max_j |\hat{m}_j - \varepsilon_j \mu_{\pi(j)}| > \delta') \le 6m\exp(-n\delta'^2/(3528L^{12}))$ for every $\delta' < \delta$, and letting $\delta' \uparrow \delta$ gives the bound with 3600. ***Step 4 (perturbation).*** With $|R_j| \le \eta \le 1$ the data are bounded by $L + 1$ and the raw moments shift by at most $3\eta(L+1)^2$; under the stated condition this shifts skewness by at most $42(L+1)^3 \cdot 3\eta(L+1)^2 \le \delta/2$, and Step 3 at margin $\delta/2$ covers the rest. □

F.1 The certificate

For each coordinate j and $r \in \{1, 2, 3\}$, the empirical Bernstein inequality (Maurer & Pontil, 2009, Theorem 4) gives an interval $[\underline{A}_{rj}, \overline{A}_{rj}]$ containing the population raw moment with probability at least $1 - \alpha/(3m)$. Outward-rounded interval arithmetic then gives an interval for the third central moment $A_3 - 3A_1A_2 + 2A_1^3$ (the sign decision, no division needed since the standardized variance is positive) and for each ordering inequality $|\kappa_j| > |\kappa_k|$ written as a polynomial inequality by cross-multiplication. On the event that all $3m$ intervals contain their targets, which has probability at least $1 - \alpha$ by the union bound, every certified sign or ordering equals the population rule applied to the representation being evaluated; otherwise the procedure abstains. For two independent runs whose population representations lie in the same B_m-orbit, so that they share the same canonical population rule, the union bound gives a $1 - 2\alpha$ statement that both certified decisions agree with that common rule. The certificate controls moment-sampling error only, and it requires a known bounded range for the statistics.

G Proofs for Section 6

Decomposition. $\varphi \in V^A(\Pi, \mathcal{M})$ iff its expectation agrees on every equivalent pair, and equivalent pairs lie in a common class, so $V^A(\Pi, \mathcal{M}) = \bigcap_\alpha V^A(\Pi, C_\alpha)$; intersect with the invariants of the global symmetries. Any global symmetry preserves observational equivalence, hence maps each class to itself and is a symmetry of each class; therefore $\mathrm{Inv}(\mathrm{Aut}(\mathcal{Z})_{\Pi, C_\alpha}) \subseteq \mathrm{Inv}(\mathrm{Aut}(\mathcal{Z})_{\Pi, \mathcal{M}})$ and $\bigcap_\alpha V^B(\Pi, C_\alpha) \subseteq V^B(\Pi, \mathcal{M})$. □

Proof of Theorem 10. Witnesses: for equivalent $J, J' \in \mathcal{M}_T$ and any relabeling ψ with $J' = \psi \cdot J$, Steps 1 to 3 of the proof of Lemma 13 give $T^{J'} \circ \psi = A \cdot T^J$ with $A \in B_m$; substituting $T^J = g_J \cdot T$ and $T^{J'} = g_{J'} \cdot T$ gives $T \circ \psi = (g_{J'}^{-1} A g_J) \cdot T$, a witness for the shared chart T. Realizability: for $g \in B_m$ and $\psi_g := T^{-1} \circ g \circ T$, the transported chart of J is $T^J \circ \psi_g^{-1} = (g_J g^{-1}) \cdot T$, again B_m-related to T, so $\psi_g \cdot J \in \mathcal{M}_T$; the observed law is preserved as in Lemma 14(c). Theorem 2, whose proof uses the class-local scope only through the quantifiers in (i) and (ii), applies verbatim on $\mathcal{M}_T$. □

Proof of Theorem 11. Let $T := T^I$ and $H := \prod_i \mathrm{Homeo}_+(\mathbb{R})$. For $h \in H$, $\psi_h := T^{-1} \circ h \circ T$ has coordinates $t_i^{-1}(h_i(t_i(z_i)))$, a coordinatewise strictly monotone bijection, so by closure $\psi_h \cdot J \in \mathcal{M}$ for every J, and the pushforward construction preserves the observed law: $\psi_h \in \mathrm{Aut}(\mathcal{Z})_{\Pi,\mathcal{M}}$. Since ψ_h preserves observational equivalence it maps the class C of I onto itself, so also $\psi_h \in \mathrm{Aut}(\mathcal{Z})_{\Pi,C}$, and the argument below applies verbatim to $\varphi \in V^B(\Pi, C)$. Let $\varphi = k \circ T \in V^B(\Pi, \mathcal{M})$. Invariance under ψ_h gives $k(h(y)) = k(y)$ for μ-a.e. y, $\mu := P_Z^I \circ T^{-1} \sim \lambda^m$, hence λ^m-a.e. For each coordinate i and rational q take h the translation by q in coordinate i; the exceptional null sets over all (i, q) have null countable union N. On N^c, by Fubini, for λ^{m-1}-a.e. value of the other coordinates the slice of k is invariant under rational translations λ^1-a.e., and rational translations act ergodically on $(\mathbb{R}, \lambda^1)$, so each slice is a.e. constant; iterating over coordinates, k is λ^m-a.e. constant, and pulling back by T, φ is P_Z^I-a.e. constant. □

H Other families: content–style and CITRIS

For content–style models with augmentations, ingredient (i) is the block-identifiability theorem of von Kügelgen et al. (2021), which gives witnesses acting on the content block by an invertible map; ingredient (ii) is a closure property of the class that must be checked. Both cases below are interface statements, not re-derivations of the cited theorems.

H.1 Content–style

Let $Z = (Z_C, Z_S)$ on $\mathcal{Z}_C \times \mathcal{Z}_S$ and let the observation model preserve content and vary style, as in the augmentation model of von Kügelgen et al. (2021). Their block-identifiability theorem is ingredient (i) with chart the content block and group the invertible maps of $\mathcal{Z}_C$ realized between equivalent instances; call the subgroup actually realized inside a class C the content stabilizer Stab. If the class realizes every relabeling of the style block (so that no function depending on z_S is invariant) and is closed under Stab (ingredient (ii)), the argument of Theorem 2, applied with the content block as chart on the quotient by style, gives $V^B(\Pi, C) = \{h(z_C) : h \text{ bounded Borel},\ h \circ \psi = h\ \forall \psi \in \mathrm{Stab}\}$, and if Stab acts transitively on $\mathcal{Z}_C$ this is $\mathbb{R} \cdot \mathbf{1}$ by the argument of Theorem 3. Which subgroup is realized depends on the class; without an anchor on content the full invertible group is available and the content collapses, with an anchor (a normalization on the content block) a proper subgroup remains. We do not claim more than this interface statement here.

H.2 CITRIS, as a sketch

Let $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_K$ with per-factor normalization and a temporal observation model in which a regime variable r_t selects the factor intervened on at step t (Lippe et al., 2022). Under three conditions that we state and do not prove here, (G1) per-factor completeness: for each factor k the conditional law of Z_k under $\{r_t = k\}$ satisfies (A1) to (A6); (G2) cross-factor decoupling under interventions: intervening on factor j disconnects its mechanism from the others, $\sigma(Z_t^k) \perp \sigma(Z_{t+1}^j \mid r_t = j)$ for $j \neq k$; and (G3) a block-factored minimal sufficient statistic $T = (T_1(z_1), \ldots, T_K(z_K))$ with $\sigma(T_k) = \sigma(Z_k)$; the equivalence class factors as a product of per-factor classes, Theorem 5 applies per factor, and the residual group is (permutations of isomorphic factors) $\ltimes \prod_k \mathrm{Stab}(\kappa_k)$, so the identified content is the bounded functions of T invariant under that group, consistent with block identifiability up to minimal causal variables. A complete proof would define the regime σ-algebra and the restricted per-factor operators, prove the product decomposition from (G1) and (G2), and verify (G3) for the exponential families used; we have not done this and present the case as a sketch.

I EXPERIMENTAL DETAILS

Statistic-level illustration (Figure 1). $m = 3$; T sampled from standardized Gamma laws with heterogeneous shape parameters so that skewnesses differ across coordinates; nine test functions in three groups (B_m-invariant, S_m-invariant only, not invariant); deviation $\max_g \mathbb{E}|h(g \cdot T) - h(T)|$ over 30 random coordinatewise monotone reparameterizations (no normalization), all 48 elements of B_3 (standard), or all 6 permutations (skewness normalization with equal targets); 10^5 samples; deviations below 0.01 are treated as zero.

Fitted iVAE (Gaussian family). True $z \mid u$ conditionally factorial Gaussian location-scale over $u \in \{1, \ldots, 6\}$ with per-(u, i) parameters chosen so the marginal skewnesses are $(1.063, 0.504)$ for $m = 2$ (two statistics per coordinate; variability rank $4/4$, smallest singular value 0.21); fixed random three-layer LeakyReLU decoder $2 \to 8 \to 8 \to 10$ with Gaussian noise $\sigma = 0.05$. Model: Gaussian conditional prior tables over u, MLP encoder from $(x, \mathrm{onehot}(u))$ to (μ, σ), MLP decoder, ELBO, Adam, 12,000 steps with cosine decay; validation ELBO on 4,000 fresh points logged every 1,000 steps; metrics on a fixed held-out set of 8,000 points. Three nested master datasets ($500 \subset 2000 \subset 8000$ prefixes, plus $n \in \{50, 125\}$), eight optimizer seeds per cell, 120 fits. Aligned error: per-coordinate mean squared error after standardization and the best signed permutation; unaligned: identity. Cell means of aligned error $0.309, 0.189, 0.238, 0.274, 0.197$ at $n = 50, 125, 500, 2000, 8000$; best restarts $0.061, 0.039, 0.040, 0.097, 0.047$; ELBO-selected restarts $0.203, 0.205, 0.040, 0.250, 0.303$. Exact-posterior floor for this metric under the true model, by two-dimensional quadrature: 0.0078. An $m = 3$ run with eight auxiliary values (variability rank $6/6$, smallest singular value 0.274) gives aligned error 0.676 ± 0.139 and correlation 0.662: recovery is weaker at $m = 3$ for the same estimator-class reason. Table 3: functionals of the standardized fitted latents; cross-restart standard deviation divided by the functional's standard deviation across observations; raw standard deviations for the skewness experiment are $0.078/0.066 \to 0.011/0.033$; modal-match consistency uses $|\mathrm{corr}|$-based coordinate matching with sign recorded separately and $\mathrm{sign}(0) \mapsto +1$; pairwise cross-restart $|\mathrm{corr}|$ of coordinate one improves from 0.764 to 0.909 (worst pair 0.164 to 0.639).

Scalar-statistic family. $p(z_i \mid u) \propto \exp(\lambda_{u,i} \tanh z_i)\varphi(z_i)$ with φ the standard normal density, $u \in \{1, \ldots, 6\}$, variability rank $2/2$ (smallest singular value 0.67), skewnesses of the statistics $(-0.24, -0.54)$; fitted with the matching prior (learnable natural parameters; Gauss–Hermite log-partition); $m = 2$, $n \in \{2000, 8000\}$, eight seeds. Aligned error $0.200/0.186$ (mean), best restart $0.044/0.029$ (correlation 0.98), unaligned $2.35/2.21$; best-aligning element not the identity in $16/16$ fits.

Certificate runs. Exact-up-to-B_m representations of the scalar-statistic family with a random $g_0 \in B_2$ per run, $n \in \{250, 1000, 4000, 16000, 64000, 256000\}$, 200 runs in total at $\alpha = 0.05$: decisions correct at every n; abstention through $n = 16{,}000$; certification of every decision at $n = 64{,}000$ and 256,000. The 16 fitted models of the scalar-statistic family: 0 certified; orientations disagree with the truth in 16 of 16. No false certification in any of the 216 runs.