Beyond ICA: Identifiability by Symmetry Breaking

Pengzhou Wu
SANKEN, The University of Osaka / RIKEN AIP
pengzhou.sk@osaka-u.ac.jp

Abstract

We prove the population-level identifiability of deep generative models (DGMs) with piecewise-affine (PWA) decoders and Gaussian mixture model (GMM) priors, in a purely unsupervised setting. We introduce three algebraic contrast principles for symmetry breaking: *domain contrast*, which trivializes the mixture symmetry group, a label-free counterpart of the sufficient-variability assumption of nonlinear ICA; *mechanism contrast*, which ensures every decoder branch gets an observational boundary across which it toggles alone; and *interaction contrast*, which forbids parameter conspiracies between latent components and decoder branches. Together they exploit the interplay between the discrete combinatorics of the PWA map and the continuous symmetry structure of the latent GMM. Continuity is replaced by algebraic symmetry conditions; injectivity is decoupled from structural identification and required only for pointwise inversion. Our results form a hierarchy: from *law identifiability* (LID; latent distribution up to a global affine map) through *map identifiability* (MID; decoder up to the same map) to *posterior* and *pointwise identifiability*. The ICA-form ambiguity emerges under conditions on diagonal component covariances. The interaction contrast is the only one of our contrasts required on the estimator side.

To our knowledge this is the first to develop algebraic symmetry-breaking into an identifiability engine for a concrete nonlinear model class, the first to admit discontinuous decoders, and the first to handle fully non-injective decoders, where every observation admits multiple latent codes.

1 Introduction

Identifiability is the foundation of trustworthy DGM. It is a population notion: the observable law (probability distribution)[1] is taken as known exactly. A deep latent variable model $X = f(Z)$ is *identifiable* if the observable law of X pins down both the map (decoder/generator) f and the latent law of Z, up to a well-defined ambiguity class. Without it, two trainings on the same data can converge to entirely different mechanisms, and downstream tasks lose any guarantee of being well-posed. Identifiability is the precondition for robust, interpretable, and disentangled representations [30], with deep classical roots in finite mixtures [2, 43, 48], tensor decompositions [27], and linear ICA [8, 9, 39]. The nonlinear, high-capacity setting of modern generators makes the ambiguity class far larger than these classical settings could admit [22].

Domain contrast in nonlinear ICA breaks ambiguity—but only under *observed* domain labels. The dominant remedy treats data as drawn from several *domains*—variations in the problem setting (gender, temperature, time, location)—each shaping the latent law inside a parametric family, and lets the diversity of domain parameters break the symmetries that otherwise plague the model: in time-contrastive learning (TCL) [21] domains are time segments; in identifiable variational autoencoder (iVAE) [25], observed auxiliaries. In both, the domain label is observed. Real data often arrives without labels, so we model the latent law as a GMM—each component being one domain. Our domain contrast distils the algebraic essence of this diversity and weakens it to unlabelled domains. Going from observed to unobserved domains is qualitatively harder: domain contrast alone no longer pins

[1] A mapping between our terminology (*law*, *map*, *domain*, *contrast*, etc.) and formal objects is given in Table 3 (App. B).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

down spurious solutions. Unsupervised identifiability is nonetheless available under continuity and smoothness, via structural sparsity [51] or restricted function classes [6]. These are assumptions we do not require.

Mechanism contrast turns observational boundaries into a resource. We add a complementary symmetry-breaking lever from the generator itself. ReLU networks are continuous PWA maps [33] and achieve almost-optimal convergence rates on discontinuous target functions [23]. A PWA generator partitions the latent space into branches with distinct affine rules—discrete *mechanisms*—and *each domain is acted on by every mechanism whose branch domain it intersects*. The algebraic resource is the boundary structure: when branches do not all share their *observational* boundaries, each toggle localises distinct constraints that no smooth global generator could ever produce. Discontinuity of f is fully allowed by our theory but is not what does the work—a continuous PWA already exhibits the boundary structure we exploit.

Algebraic symmetry collapse replaces continuity and smoothness. Existing nonlinear-ICA proofs are differential at heart—log-density and Jacobian comparisons [16, 17, 21, 25]—silent on discontinuous maps and boundary structure, and structurally requiring smoothness. We replace differential techniques with an *algebraic* engine: collisions between two candidate models are read as nontrivial elements of an affine symmetry group of the latent law, and assumptions that trivialise this group force the two models to coincide. Three contrast principles—one each for map, law, and their interaction—power the engine (Table 1; formal in §3). The combination keeps the merits of both worlds: the affine link is built from map-side boundary information, so domain labels can remain latent; it trades one assumption profile for another rather than dominating. Symmetry arguments have emerged in identifiability since the 1950s [20]; classical models had small, obvious symmetry groups [3], so a general algebraic engine is needed only in the nonlinear, high-capacity regime.

Generative modelling is not signal processing. Classical ICA inherits its goals from signal processing: invert the source pointwise. This requires injectivity, a constraint both restrictive in theory and almost never enforced in mainstream nonlinear-ICA learning procedures [16, 21, 25], even when their theorems demand it. Deep generative modelling cares about a different object: the *mechanism* (the decoder) and the *posterior* (the representation). Neither needs pointwise inversion. Our hierarchy (law $\to$ map $\to$ posterior $\to$ pointwise; Table 2) delivers the first three *without* injectivity; pointwise recovery is an optional upgrade. *No* form of latent independence is needed for the affine link: Theorems 5 and 8 hold for any component covariances. Only the ICA-form refinement (Cor. 11) requires them to be diagonal, i.e., conditional independence of coordinates given the component.

Table 1: The three contrast principles powering the engine. Domain contrast (MST) refines the classical observed-domain assumption; mechanism contrast (SB)/(USB) is new and exploits PWA branch boundaries; interaction contrast (AD) is the parameter-level analogue of Independent Causal Mechanisms.

Contrast	**Asn.**	**Constrains**	**Informal content**
domain	(MST)	latent law	no affine change of latent coordinates maps the mixture onto itself (replaces TCL/iVAE's sufficient variability)
mech., existential	(SB)	map	an observational boundary across which a solo branch toggles
mech., universal	(USB)	map	every branch gets such a boundary
interaction	(AD)	map–law coupling	map and latent parameters do not conspire

Contributions.

1. *Mechanism and interaction contrast* as principles complementary to domain contrast. *Takeaway:* branch boundaries—and the discontinuities they may carry—are a resource, not an obstruction.
2. *A group-theoretic identifiability engine.* Continuity and differentiability are replaced by triviality of an affine symmetry group, freeing pushforward identifiability from the smoothness restrictions of the differential proof tradition. *Takeaway:* a single algebraic principle (symmetry collapse) carries the analysis from law identifiability through ICA-form recovery, and shows iVAE's sufficient variability as symmetry breaking.
3. *A hierarchy of identifiability* (law $\to$ map $\to$ posterior $\to$ pointwise), with conditions modular along the way. *Takeaway:* identifiable representation learning is achievable without injectivity and without any latent independence assumption.
4. *Truth-sided assumptions.* (AD) is the only one of our contrasts required on the estimator side. The other contrasts live on the truth side. *Takeaway:* learning methods need not bake every theoretical condition into architecture, saving compute and admitting flexibility.

2 Related Work

A comprehensive comparison along all assumption axes is in Table 4 (App. C); detailed technical comparisons with iVAE [25], Kivva et al. [26], and boundary-richness vs. injectivity appear in Appendix U. Here we highlight the lineages most relevant to our positioning.

Nonlinear ICA. TCL [21] and iVAE [25] identify smooth invertible mixings of independent latents whose marginals vary across observed segments or auxiliaries. In effect, our (MST) weakens TCL/iVAE's sufficient-variability assumption: it needs no labels, and generically as few as two components, where TCL/iVAE needs more than $2d$ for Gaussian latents of dimension d (Remarks 4.6 and 4.7). With a volume-preserving mixing, Yang et al. [49] also need only two classes, but observed ones. Multi-view nonlinear ICA [16] uses several full mixings of a shared independent latent under a sufficiently-distinct-views condition. Independent mechanism analysis (IMA) [17] postulates Jacobian-column orthogonality of a smooth invertible mixing and is implemented through normalising flows that are invertible by construction. The mechanism-sparsity line [4, 5, 28, 32, 50, 51] places interaction conditions on smooth invertible decoders. All are differential at heart and assume smooth, typically invertible generators. Our (USB) is a PWA-geometric analogue of their structural-sparsity conditions (App. U).

Relaxing injectivity and continuity. To our knowledge, only three works in this lineage (partially) relax injectivity. Kivva et al. [26] require injectivity *on an f-image region* for LID and global injectivity for MID; Matthes et al. [32] focus on local disentanglement under local injectivity; Roeder et al. [35] obtain linear identifiability for *discriminative* softmax-form models—a different paradigm from generative pushforwards. Continuity of f is dropped only by us; Kivva et al. [26] rely on continuity for density matching to obtain MID, though allow discontinuity for LID.

Group symmetry in identifiability and disentanglement. Higgins et al. [19] define disentanglement via a *given* direct-product decomposition of a group on world states; the latent law plays no role and there is no theorem for any concrete model class. Xi and Bloem-Reddy [46] taxonomise model indeterminacies as $\mathcal{A}(\mathcal{F}) \cap \mathcal{A}(\mathcal{P}_z)$; they require injective generators and resolve multi-domain ambiguity by *observed* domain labels, and handle the GMM analogue only by collapsing each domain to a single Gaussian (their §4.3, Thm. 4.5)—bypassing rather than trivialising the within-domain symmetry that our (MST) addresses. Concurrently, Squires and Ravikumar [41] reduce identifiability to whether an intersection of two transition sets lies in a target group, extending Xi and Bloem-Reddy [46] to stochastic mixing. For PWA–GMM, (AD) and (SB)/(USB) constrain the transitions between decoders, and (MST) trivialises the latent law's own symmetry group, not the target group $\mathrm{Aff}(\mathbb{R}^d)$ (App. U). Ahuja et al. [1] place structure on *known* latent dynamics: shared equivariances of the dynamics define the ambiguity class, with smooth invertible observation map throughout. Group ideas long appear in invariant *inference* [12, 15], starting from a *prescribed* group action.

PWA + GMM, and multi-mechanism modelling. Kivva et al. [26] opened the unsupervised PWA+GMM regime; Xu et al. [47] extend to potentially-degenerate GMMs via encoder sparsity (still injective continuous PWA); Lopez et al. [31] specialise Kivva to comparative DGMs with an *observed* contrastive split. Bayesian ICA with mixture priors [7] and VaDE, GMVAE and MFCVAE [10, 13, 24] pursued this class empirically. For its equidimensional case, our conditions are sufficient for identifiability. Generalised PCA [44] and clusterwise linear regression [18] are remote relatives—single-condition projective settings without unobserved domain labels or branch analysis. Modelling generating processes via multiple temporal or spatial mechanisms is classically well-motivated: temporal exemplars include switching/hybrid systems [34], recurrent switching linear dynamics [29], and PWA neuroscience dynamics [11]; spatial exemplars include free-form-deformation and PWA image registration [14, 36, 40, 42]. None addresses identifiability; we bring algebraic identifiability to this PWA modelling tradition.

3 Setup and preparations

A latent random variable Z follows a GMM and is observed through an unknown PWA map f; that is, $X = f(Z)$.[2] Both the map and the latent law are hidden: only the data of X is available. We ask whether Z and f can be (partially) recovered from this data.

[2] Additive independent noise ε with non-vanishing characteristic function can be handled via deconvolution; see e.g., [25].

3.1 PWA maps and Gaussian mixture models

Definition 3.1 (Irredundant GMM)**.** A random variable Y is an *irredundant GMM* with K components if its density is $\sum_{k=1}^{K} w_k \, \varphi_{\mu_k,\Sigma_k}$, with φ_{μ_k,Σ_k} the density of $\mathcal{N}(\mu_k, \Sigma_k)$, where all weights are positive, all covariances are positive definite, and parameters $(\mu_k, \Sigma_k) \neq (\mu_\ell, \Sigma_\ell)$ for $k \neq \ell$.

Definition 3.2 (Reduced PWA map)**.** A map $h\colon \mathbb{R}^d \to \mathbb{R}^d$ is a *reduced PWA map* if it is specified by a finite index set I and, for each $i \in I$, an invertible affine rule (A_i, b_i) together with a branch domain $P_i \subset \mathbb{R}^d$[3], such that each P_i is a finite union of full-dimensional polyhedra with pairwise disjoint interiors; the branch domains $\{P_i\}_{i\in I}$ partition $\mathbb{R}^d$ up to measure-zero boundaries; and the rule map $i \mapsto (A_i, b_i)$ is injective.

Reducedness prevents duplicate bookkeeping: knowing the local affine rule determines the branch index. This is not a restriction in practice—any PWA map can be converted to a reduced one by merging branches with identical affine rules (their unions are always finite unions of full-dimensional polyhedra). A single branch index may therefore be active on several disconnected regions of the observational space; this causes no issue in the analysis below.

Definition 3.3 (PWA equality)**.** Two reduced PWA maps f and g are *equal*, written $f = g$, if there exists a bijection $\pi\colon I_g \to I_f$ such that matched branches have identical affine rules ($B_j = A_{\pi(j)}$ and $c_j = b_{\pi(j)}$) and their domains coincide up to measure-zero boundaries.

Convention. Throughout, *PWA map* and *GMM* refer to reduced PWA maps and irredundant GMMs, respectively. Equalities of branch domains and chambers (§3.3) hold up to null sets.

3.2 Identifiability

We fix an arbitrary *truth* pair (f, Z) and an *estimator* pair (g, Z'), both PWA–GMM pairs in $\mathbb{R}^d$. We observe $X = f(Z)$. The central assumption throughout is the *distributional equality*

$$f(Z) \sim g(Z'), \tag{1}$$

meaning the two models produce the same observable distribution. We consider a hierarchy of identifiability notions, from weakest to strongest, summarised in Table 2.

Table 2: Identifiability hierarchy. (AD) on both sides is always assumed; (SB), (USB) and (MST) are on the truth side only. '+' means: in addition to all assumptions in the rows above. Under (DD) on both sides (conditional independence required), α takes the ICA form $P\Lambda z + c$ at every level (Cor. 11).

Level	Assns.	Conclusion	Ref.	Meaning
LID	(AD) + (SB)	$\alpha(Z') \sim Z$	Thm 5	Latent distribution recovered up to a global affine reparametrization.
MID	+ (USB) + (MST)	$f = g \circ \alpha^{-1}$	Thm 8	Generative mechanism recovered up to α.
Post.	(directly from MID; no additional assns)	posteriors linked by α	Prop 9	Posterior inference translates between the two models.
Point.	+ f injective	$z = \alpha(\hat{z})$	Prop 10	True source recovered pointwise (classical ICA goal).

Each level builds on the previous. Under (DD) (conditional independence of coordinates given the component), the affine link α is further refined to the ICA form $\alpha(z) = P\Lambda z + c$ (Cor. 11); this is not a new rung in the hierarchy, but a cross-cutting refinement that sharpens α at every level.

3.3 Chambers and the pushforward density

Identifiability relies on observational information. To analyze (1), we decompose the *observational space* into regions, i.e., *chambers*, where a map acts with constant branch structure. This converts the global PWA pushforward equality into a *local* matching of finite Gaussian mixtures.

Definition 3.4 (Chambers of a single map)**.** For a PWA map h with branch index set I, the *active set* at $x \in \mathbb{R}^d$ is $\mathcal{I}^h(x) := \{i \in I : x \in h(P_i)\}$. A *chamber* of h is a maximal open connected subset of $\mathbb{R}^d$ on which $\mathcal{I}^h(\cdot)$ is constant.

[3]The *branch domain* P_i is a piece of the map's *input* space $\mathbb{R}^d$; it should not be confused with a "domain" in the multi-domain sense (a mixture component / environment).

Informally, the active set lists the branches that could have produced the observation x, and a chamber keeps it constant. For example, $h(z) = |z|$ on $\mathbb{R}$ has active set $\{1, 2\}$ on $x > 0$. See Appendix Figure 8 for a 4-branch PWA map with 4 chambers. In general, the chamber structures of two different maps f and g may not coincide: one map's chamber boundary may fall in the interior of another map's chamber. We therefore refine the partition to the common one.

Definition 3.5 (Fine chambers)**.** Given two PWA maps f, g, the *fine chambers* are the maximal open connected subsets of $\mathbb{R}^d$ on which both $\mathcal{I}^f(\cdot)$ and $\mathcal{I}^g(\cdot)$ are simultaneously constant.

Informally, they overlay the chamber boundaries of both maps. On each fine chamber C, we write $\mathcal{I}^f(C)$ and $\mathcal{I}^g(C)$ for the constant active sets. The *pushforward* of the k-th latent component under truth branch i is the Gaussian $\mathcal{N}(\theta_{i,k}^{(f)})$ with parameters $\theta_{i,k}^{(f)} := \big(A_i\mu_k + b_i,\ A_i\Sigma_k A_i^\top\big)$, and analogously $\theta_{j,\ell}^{(g)}$ for (g, Z'). Inside C, the distributional equality (1) becomes a pointwise identity between two finite Gaussian mixtures:

$$\sum_{i\in\mathcal{I}^f(C)} \sum_{k=1}^{K} w_k\, \varphi_{\theta_{i,k}^{(f)}}(x) = \sum_{j\in\mathcal{I}^g(C)} \sum_{\ell=1}^{K'} w'_\ell\, \varphi_{\theta_{j,\ell}^{(g)}}(x) \tag{2}$$

for all $x \in C$. Crucially, each Gaussian kernel $\varphi_{\theta_{i,k}^{(f)}}$ is defined globally on $\mathbb{R}^d$ (its parameters come from the branch rule (A_i, b_i) and the component parameters, not from C). Extending both sides of (2) by analytic continuation from the chamber C produces two globally-defined analytic functions F_C and G_C that agree everywhere on $\mathbb{R}^d$. We stress that this is an analytic extension of the *chamber-mixture function*, not of the true pushforward density—the latter is given by different active sets on different chambers. In particular, different fine chambers generally yield different analytic extensions, and we never equate extensions across chambers. This local–global interplay is a recurring theme of our results.

We can already draw a nontrivial conclusion without any further assumptions.

Lemma 1 (Active-branch-count equality)**.** *Under* (1)*, the number of active branches is the same on both sides of every fine chamber:* $|\mathcal{I}^f(C)| = |\mathcal{I}^g(C)|$.

In other words, the total "mass budget" on each fine chamber is determined by the number of active branches, and this number must match on both sides.

4 Theoretical results

Full proofs of all results appear in Appendix A; proof sketches for the key results are given below. Illustrated examples are given in Appendix K and L.

Proof strategy. Our whole proof strategy has an *algebraic* flavour and is built around *bijections* between the truth and estimator, at progressively higher levels. (AD) gives a *branch-component-pair bijection* Ψ. (SB) restricts Ψ to a single boundary and reads off a component-level affine link α: it "matches"–via the induced *component bijection*–every estimator component to a truth component *and* "stitches" the latent components together by this same α, yielding LID. (USB) gives a *branch bijection*: it "matches" every estimator branch to a truth branch via its own simple boundary. (MST) upgrades the branch bijection to a global branch-level affine α: it "stitches" across all branches by trivialising the mixture symmetry group, completing MID.

4.1 Pushforward atom distinctness and consequences

We now introduce the main genericity condition on map–latent pairs.

Condition (AD) (Atom distinctness).[4] Call the pushforward Gaussian $\mathcal{N}(\theta_{j,\ell})$ of component ℓ through branch j an *atom*. A PWA-GMM pair (h, Y) satisfies (AD) if its atoms are pairwise distinct, i.e., the map from branch–component indices $(j, \ell) \mapsto$ pushforward parameters $\theta_{j,\ell}$ is injective.

This is the *interaction contrast* of Table 1: it formalises the principle that the map rules and the latent parameters should not conspire to produce collisions in the observational space. Collisions within the same branch are already excluded by invertibility of the branch rule composed with irredundancy of the GMM; (AD) additionally forbids cross-branch collisions. It decomposes into two sub-conditions:

[4] Called parameter injectivity, (PI), in the preprint version of this paper.

1. Same-component (AD): no collision $\theta_{j,\ell} = \theta_{m,\ell}$. A violation means the *branch-difference map* $h_m^{-1} \circ h_j$ is a *symmetry* of component $\mathcal{N}_\ell$.
2. Cross-component (AD): no collision $\theta_{j,\ell} = \theta_{m,n}$ with $\ell \neq n$. Such a collision forces the affine difference map $h_m^{-1} \circ h_j$ to *transport* component $\mathcal{N}_\ell$ exactly onto $\mathcal{N}_n$.

Both are *generic*, i.e., hold almost surely (dimension counting in Appendix F). (AD) is the parameter-level analogue of Independent Causal Mechanisms [37]: GMM parameters are the mechanism for how Z is generated, and affine branch rules are the mechanism for how X is generated from Z; (AD) forbids conspiracy between them at the parameter level.

Chamber-wise (AD)—atom distinctness restricted to active pairs per chamber—is weaker than the stated global version; the glueing step in Lemma 2 requires the full global condition (see Appendix E for the distinction and a counterexample).

We first record a standard but central fact: distinct non-degenerate Gaussian densities are linearly independent as functions (see Appendix D for the precise statement). We denote this below as Lemma 18 and refer to it as the *Gaussian-independence lemma*.

Remark 4.1 (Equality on open sets implies equality everywhere). If two finite Gaussian mixtures agree on any open set $U \subseteq \mathbb{R}^d$, they agree on all of $\mathbb{R}^d$: both are real-analytic, and $\mathbb{R}^d$ is connected. The Gaussian-independence lemma then applies to their difference: *parameters appearing on one side must also appear on the other, and matched parameters carry equal coefficients.* We will use this observation repeatedly when analysing (2) on each fine chamber.

Lemma 2 (Global parameter bijection under (AD))**.** *Assume* (1) *and that both* (f, Z) *and* (g, Z') *satisfy (AD). Then there exists a unique bijection*

$$\Psi \colon I_g \times [K'] \;\longleftrightarrow\; I_f \times [K] \tag{3}$$

such that $\theta^{(g)}_{j,\ell} = \theta^{(f)}_{i,k}$ *and* $w'_\ell = w_k$ *whenever* $\Psi(j, \ell) = (i, k)$*. On each fine chamber* C*,* Ψ *restricts to a bijection between the active pairs on each side.*

Proof sketch. On each fine chamber C, both sides of (2) are finite GMMs agreeing on an open set; Remark 4.1 and (AD) yield a bijection Ψ_C matching active branch-component pairs by parameter equality. Uniqueness of each match (by (AD) on both sides) forces all chamber-wise bijections to agree: if (j, ℓ) and (i, k) are matched on some C, then $\theta^{(g)}_{j,\ell} = \theta^{(f)}_{i,k}$ is a global identity, so $\Psi_{C'}(j, \ell) = (i, k)$ for any other chamber C' containing branch j. This is the step that requires *global* rather than merely chamber-wise (AD) (Remark E.1). □

Lemma 1 (no (AD)) and Lemma 2 (with (AD)) form the *foundational machinery* reused throughout this paper. The former controls the branch-count budget; the latter turns density matching into a discrete combinatorial bijection. Both play the role of *matching* the estimator to the truth at the most basic level; all higher-level identifiability results will build on this matching. We assume double (AD) (on both sides) hereafter.

For adjacent fine chambers (C, C') (two whose closures share a $(d-1)$-dimensional boundary), define the *leaving count* $D_f(C, C') := |\mathcal{I}^f(C) \setminus \mathcal{I}^f(C')|$, and symmetrically for g. By swapping $C \leftrightarrow C'$, one obtains *entering counts* E_f, E_g.

Matching identifies three combinatorial invariants—the component count, the branch count, and the chamber structure (Lemmas 14, 15, and 16 in Appendix A), collected in Proposition 3.

Proposition 3 (Consequences of matching)**.** *Under* (1) *and double (AD): the component counts agree,* $K' = K =: K$*, which immediately implies* $|I_f| = |I_g|$*; at every pair of adjacent fine chambers the leaving and entering branch counts agree,* $D_f(C, C') = D_g(C, C')$ *and* $E_f(C, C') = E_g(C, C')$*; and consequently* f*-chambers,* g*-chambers, and fine chambers all coincide.*

We write $K := K = K'$ throughout the rest of the paper. It is worth stressing that $|I_f| = |I_g|$ does *not* follow from active-count equality (Lemma 1) alone: chambers generally share branches, and before (AD), we have no reason to expect the sharing patterns of the two maps to match. It is the global bijection Ψ (Lemma 2) that forces $|I_f| = |I_g|$, and then the boundary counting forces the sharing patterns themselves to agree.

We begin with the simplest but already nontrivial identifiability setting: the observable law is itself a GMM. This corresponds to the true map being the identity, $f = \mathrm{Id}$.

Theorem 4 (From PWA to affine: the symmetry-rigidity)**.** *Let g be a PWA map, Z' a GMM, and suppose (g, Z') satisfies (AD). If $g(Z') \sim Z$, where Z is a GMM, then g is globally affine: $g(x) = Ax + b$ for some $A \in \mathrm{GL}(d),\ b \in \mathbb{R}^d$, and for all x.*

In particular, if $g(Z') \sim Z'$, g is globally affine. That is, **no** strict PWA *self-map satisfies (AD).*

Remark 4.2 (Essential role of (AD)). Theorem 4 highlights why (AD) is essential for *any* identifiability result: if a GMM p admitted a *strict* PWA self-map S (at least two branches, $S_\sharp p = p$), then for any truth map f and $Z \sim p$, setting $g := f \circ S$ and $Z' := Z$ yields $g(Z') = f(S(Z)) \sim f(Z)$. Here g is genuinely different from any affine reparametrization of f, so identifiability fails. (AD) is precisely the condition that rules out such pathologies: by Theorem 4, any PWA self-map of a GMM respecting (AD) must be globally affine.

Remark 4.3 (Hard cases). Theorem 4 also implies: under double (AD), if either f or g is globally affine, so is the other. Indeed, if f is affine, then $f(Z)$ is a GMM, and the theorem forces g to be affine. The focus in the sequel is the nontrivial case where *both* maps are *strictly PWA* (at least two branches) and the pushforward is *not* a GMM.

4.2 Law identifiability

We now tackle the first nontrivial step in the hierarchy: *law identifiability (LID)*, where we show the two latent GMMs are related by a global affine map. This requires a boundary-genericity condition—the *mechanism contrast* of Table 1.

Condition (SB) (Simple boundary). A PWA map h satisfies (SB) if either (a) h has a single chamber (i.e., h is globally affine; Lemma 13), or (b) there exist adjacent chambers (C, C') of h and a branch $i_\star$ that is the *unique* branch leaving *or* entering ($\mathcal{I}^h(C) \setminus \mathcal{I}^h(C') = \{i_\star\}$ *or* $\mathcal{I}^h(C') \setminus \mathcal{I}^h(C) = \{i_\star\}$).

(SB) exposes *observational* boundary information, which Theorem 5 employs for identification. (SB) constrains only *one direction* of the boundary crossing (leaving *or* entering); the opposite direction is unrestricted. Thus a boundary with one branch leaving and three branches entering qualifies as simple. **Note:** (SB) is satisfied even if multiple branches overlap across many boundaries, provided *at least one* directional boundary crossing changes the active set by exactly one element.

Remark 4.4 (Genericity of (SB)). The violation of (SB) for a strict PWA map is a rigid condition: *every* boundary must have at least two branches leaving together *and* at least two branches entering together. Each chamber boundary of a generic PWA map is typically traversed by a single branch, so (SB) is generic under unconstrained parametrizations (App. L.7). For continuous PWA maps, genericity is a conjecture: co-toggling can happen at shared breakpoints (Appendix L.3), but a *single* non-co-toggling boundary suffices.

Theorem 5 (Law identifiability)**.** *Assume* (1) *with (AD) on both sides and (SB) on f. Then $K' = K =: K$, and there exists an affine bijection $\alpha \in \mathrm{Aff}(\mathbb{R}^d)$ such that $\alpha(Z') \sim Z$; more precisely, there is a permutation σ of $[K]$ with $\alpha_\sharp \mathcal{N}(\mu'_\ell, \Sigma'_\ell) = \mathcal{N}(\mu_{\sigma(\ell)}, \Sigma_{\sigma(\ell)})$ and $w'_\ell = w_{\sigma(\ell)}$ for every ℓ.*

Proof sketch. Lemma 14 gives $K' = K$. The (SB) boundary yields unique toggling branches $i_\star$ (truth, WLOG leaving) and $j_\star$ (estimator, by Lemma 15). The bijection Ψ restricted to $\{j_\star\} \times [K] \leftrightarrow \{i_\star\} \times [K]$ induces a permutation σ and, via the parameter equality, determines the affine map $\alpha(z) := A_{i_\star}^{-1}(B_{j_\star} z + c_{j_\star} - b_{i_\star})$, which transports each estimator component to the corresponding truth component with matching weights. □

The link α is constructed from a specific toggling pair—it is not claimed canonical. LID says only that *some* affine link exists. What will matter for MID is *globalness*: forcing the same structural relationship across every branch, which is exactly what (MST) delivers.

4.3 Map identifiability

We now address the stronger question: must the maps f and g themselves coincide (up to the latent reparametrization α)? The key challenge is that LID establishes the existence of α via a *single* toggling boundary, but says nothing about other branches. For MID, we must ensure that *every branch* of g is related to the corresponding branch of f by the *same* affine link. This requires two additional ingredients: boundary richness (every branch is exposed at some simple boundary) and symmetry triviality—the *domain contrast* of Table 1.

Condition (USB) (Universal simple boundary). A PWA map h satisfies (USB) if either (a) h has a single chamber, or (b) for every branch $i \in I$, there exist adjacent chambers (C, C') of h such that i is the unique leaving branch or the unique entering branch.

(USB) is *universal for branches*: every branch is exposed at some simple boundary. It does not require every boundary to be simple. (USB) trivially implies (SB). Similar to (SB), (USB) is satisfied even if multiple branches overlap across many boundaries. (USB) is *strictly weaker than global injectivity*; instead of weak injectivity [26], which demands a fragile, globally unopposed 1-active chamber, (USB) relies only on a robust, local combinatorial asymmetry (Appendix L.4&L.7).

Remark 4.5 (Genericity of (USB)). A violation of (USB) requires that for some branch i, every chamber boundary of i is simultaneously shared by at least one other branch on both the leaving and entering sides. Like (SB), (USB) is therefore generic under unconstrained parametrizations.

We next introduce the symmetry condition that "locks" the identification. Given a GMM with density p, its *mixture symmetry group* is $G_{\mathrm{mix}}(p) := \{T \in \mathrm{Aff}(\mathbb{R}^d) : T_\sharp\, p = p\}$. This group captures all affine bijections that preserve the mixture law as a whole. It can include maps that *permute components of equal weight* (not just maps that fix each component individually; the distinction is important and discussed in Appendix G.3).

Condition (MST) (Mixture symmetry triviality). A GMM p satisfies (MST) if $G_{\mathrm{mix}}(p) = \{\mathrm{Id}\}$.

From the symmetry standpoint, (MST) is stronger than (AD): (AD) rules out *strict-PWA* self-maps of the mixture (Remark 4.2), while (MST) further rules out *affine* self-maps beyond the identity. (MST) is required only for MID and above; likewise $K > 1$ is needed only there. Indeed, for $K = 1$ (a single Gaussian), LID is trivial: any two non-degenerate Gaussians on $\mathbb{R}^d$ differ by an affine bijection.

(MST) is also *necessary* for MID in a structural sense; see Example 3 for a prototype ("101") construction. And this mechanism can be extended to *any* GMM and beyond (Appendix O).

Remark 4.6 ((MST) weakens iVAE's sufficient variability). iVAE's assumption (iv) is posed for exponential families; for Gaussian latents with observed labels, it implies (MST) (Proposition 30). Labels force every symmetry to fix each component; the $2d + 1$ components that (iv) requires then have diagonal precisions spanning all diagonal directions, which leaves only coordinate sign flips, and their varying means rule those out. (MST) is *strictly weaker* in two ways. With labels, $d + 1$ components with affinely independent means already give it (Lemma 23), where (iv) needs $K > 2d$. Without labels, (iv) cannot be stated, while (MST) is a property of the unlabelled mixture alone.

Remark 4.7 (Genericity of (MST)). For $K \geq 2$ components with generically chosen means, covariances, and weights, (MST) holds. A violation requires either (a) a nontrivial affine map fixing every component individually, which is ruled out by, e.g., affinely independent means (Lemma 23 in Appendix), or (b) a nontrivial affine map permuting two or more components—i.e., the components are partitioned into cycles and the map acts one-to-one within each cycle—which is ruled out by, e.g., distinct weights *or* distinct covariance determinants (Lemma 22 in Appendix). Other generic and sufficient conditions for (MST) exist; see Appendix I.3 for a summary. All the conditions are algebraically rigid and occur on a measure-zero set. See Figure 5 in Appendix K for an illustration of why case (b) matters.

Truth-sided assumptions. Before proceeding to the proofs, we record two invariance results that clarify on which side(s) each assumption needs to live. The first says that (USB) transfers to the estimator side already from double (AD)—without constructing any affine link.

Proposition 6 (((SB)) and ((USB)) inheritance)**.** *If $f(Z) \sim g(Z')$, (f, Z) and (g, Z') satisfy (AD), and f satisfies (SB) (resp. (USB)), then g satisfies (SB) (resp. (USB)).*

Moreover, (MST) transfers via affine conjugacy once an LID link is established: $G_{\mathrm{mix}}(p_{Z'}) = \alpha^{-1} G_{\mathrm{mix}}(p_Z)\alpha$, so p_Z satisfies (MST) iff $p_{Z'}$ does (Proposition 17 in Appendix A). Observational chamber structure is also preserved by any affine latent reparametrization, so (SB)/(USB) are invariant under α. In summary, (AD) is the only contrast required on the estimator side; ((SB), (USB), (MST)) are imposed on the truth side, and their propagation to the estimator is automatic.

We first establish MID in the simpler case where both maps push forward the *same* GMM law.

Lemma 7 (MID with shared latent)**.** *Let f, g be PWA maps, and let Z, Z' both follow the same GMM law p with K components. Assume* (1)*, (AD) on both sides, (USB) on f, and (MST) on p. Then $f = g$.*

Proof sketch. By (USB), each truth branch i has a simple boundary, and Lemma 15 forces a unique estimator branch $j(i)$; the composite $S_i := f_i^{-1} \circ g_{j(i)}$ then lies in $G_{\mathrm{mix}}(p)$ by the parameter match. The domain-contrast condition $G_{\mathrm{mix}}(p) = \{\mathrm{Id}\}$ ((MST)) forces $S_i = \mathrm{Id}$, so $f_i = g_{j(i)}$—simultaneously killing all branch-wise ambiguities and delivering the global affine link required for MID. Finally, injectivity of $i \mapsto j(i)$ (Proposition 6) promotes it to a bijection, and domain coincidence follows from Lemma 16 and invertibility of the branch maps. □

The general MID result follows by reducing to the shared-latent case.

Theorem 8 (Map identifiability)**.** *Assume* (1) *with (AD) on both sides, and (USB) plus (MST) on the truth side* (f, Z)*. Then there exists an affine bijection* $\alpha \in \mathrm{Aff}(\mathbb{R}^d)$ *with* $\alpha(Z') \sim Z$ *and* $f = g \circ \alpha^{-1}$*.*

Proof sketch. Since (USB) implies (SB), Theorem 5 first yields an affine α with $\alpha(Z') \sim Z$. Setting $h := g \circ \alpha^{-1}$ preserves the reduced-PWA structure, and the parameter map for (h, Z) inherits injectivity from (g, Z'), so (AD) holds for (h, Z). Lemma 7 applied to f and h (same latent Z, both satisfying the conditions) then gives $h = f$, i.e. $f = g \circ \alpha^{-1}$. □

Alternative purely algebraic routes replacing (USB) with orbit conditions appear in Appendix Q.

4.4 Posterior, pointwise, and ICA-form identifiability

MID immediately yields posterior and pointwise identifiability.

Proposition 9 (Posterior identifiability)**.** *Under the conclusions of Theorem 8, for* P_X*-almost every observation* x*,* $P_{\mathrm{est}}(\cdot \mid x) = (\alpha^{-1})_\sharp P_{\mathrm{true}}(\cdot \mid x)$*.*

Remark 4.8 (Practical significance). In Bayesian inference or downstream tasks that integrate over latent variables, the posterior $P(z \mid x)$ is the fundamental object. Proposition 9 guarantees that this posterior is identified up to α, even when the map is non-injective and the posterior has support on multiple latent points.

Proposition 10 (Pointwise identifiability)**.** *If additionally* f *is globally injective, then for every* x *in its image,* $z = \alpha(\hat{z})$ *where* $\hat{z} := g^{-1}(x)$*.*

Throughout the paper, Proposition 10 is *the only result that uses injectivity* of the map.

ICA-form ambiguity under conditional independence. Under (DD) on both sides (definition in Appendix J), the affine link α shared by all the preceding results (Theorems 5 and 8, Propositions 9–10) has the ICA form. Alternative sufficient conditions, including the route via iVAE assumption (iv), are in Appendix J.

Corollary 11 (PWA-ICA identifiability)**.** *Under the assumptions of Theorem 5 or 8, if both* Z *and* Z' *additionally satisfy (DD), then* $\alpha(z) = P\Lambda z + c$ *for a permutation matrix* P*, a diagonal invertible matrix* Λ*, and* $c \in \mathbb{R}^d$*.*

This is *the only result that uses conditional independence*; nowhere else throughout the paper latent independence—conditional or not—appears. iVAE's conditionally factorial prior is (DD) in our setting: it buys the permutation-scaling form of α, not identifiability itself.

4.5 Stability and a numerical illustration

Theorem 8 assumes exactly equal laws. Its conditions being open in the parameter space, App. R asks them to hold with a margin $c > 0$ within a complexity budget κ, which defines a class $\mathcal{F}_\kappa$ of pairs with whitened latent laws and a ball B_R. Here W_1 is the Wasserstein-1 distance.

Proposition 12 (Stability)**.** *There is a nondecreasing* $\omega = \omega_\kappa$*, with* $\omega(\varepsilon) \to 0$ *as* $\varepsilon \to 0$*, such that for all* $(f, Z), (g, Z') \in \mathcal{F}_\kappa$*: if* $W_1(f_\sharp p_Z, g_\sharp p_{Z'}) \le \varepsilon$*, then a single orthogonal* O *satisfies* $\|f - g \circ O^{-1}\|_{L^1(B_R)} \le \omega(\varepsilon)$ *and* $W_1(p_Z, O_\sharp p_{Z'}) \le \omega(\varepsilon)$*.*

Identifiability thus holds approximately: within $\mathcal{F}_\kappa$, close observed laws force close decoders and latent laws, up to one orthogonal map. On a one-dimensional family varying only the (MST) margin, the sample size a likelihood-ratio test needs to tell the true decoder from a competitor grows as the inverse square of that margin, from 29 at unwhitened margin 0.64 to 29 211 at 0.02 (App. S).

Checking the conditions in a trained model. No training objective can supply what the true pair lacks, so the data-generating process comes first, with domain knowledge as a guide. On a fitted

model, (AD) is the only contrast required of the estimator (Prop. 6), (MST) has testable sufficient conditions (Lemmas 22, 23), (USB) needs one exposing boundary per branch, probed by sampling boundary crossings, and each returns a margin (App. R). The standard ELBO enforces none of them, and adding the margins as an inductive bias may help learning; we leave this to future work.

5 Conclusion and prospects

We have established an identifiability hierarchy for DGMs with PWA decoders and GMM priors, proceeding from LID through MID to posterior and pointwise identification, and finally to ICA-form recovery. Three algebraic contrast principles—domain ((MST)), mechanism ((USB)), and interaction ((AD))—power an engine that replaces both the topological continuity glue of prior work and the observed labels of the TCL/iVAE line, whose sufficient variability our (MST) weakens. Structural identifiability (recovering the decoder and the posterior) is separated from pointwise inversion, which requires injectivity.

Genericity. (AD) and (MST) are generic: violated only on a measure-zero algebraic subset of the relevant parameter space (Appendix F, I; Remark 4.7). (SB)/(USB) are generic under unconstrained PWA parametrizations (Remarks 4.4, 4.5), and conjectured so under a continuity constraint (App. L.7). Concretely: under an unconstrained parametrization, if the branch matrices and biases of PWA map f and parameters of GMM p_Z are drawn randomly and independently, identifiability holds almost surely. To our knowledge, this is the first setting in which *all* identifiability assumptions are simultaneously generic, without requiring observed domain labels, injectivity, or continuity.

Necessity and prospects. (SB) is necessary for this class: Appendix K.1 gives a fold example, which can be generalized to a family of examples on $\mathbb{R}^d$, where (AD) and (MST) hold yet LID breaks. (MST) is necessary: Appendix O proves that if a GMM admits any nontrivial affine self-symmetry, it admits a finite-order one; such torsion symmetries can be localised into strict PWA self-maps breaking identifiability. Analogous torsion-based necessity phenomena hold for broad exponential-family and location-scale mixtures (Table 5). We conjecture that the gap between chamber-wise and global (AD) is structurally necessary in the non-injective setting, and that investigating it leads toward a full necessity result for (AD). More broadly, Gaussian mixtures and PWA maps are the first testbed for a *general algebraic identifiability theory*: one can replace the affine group with the natural symmetry group of a general component family F (e.g., any exponential family), study mixtures from F and decoder branches from that F-symmetry group, and ask whether analogues of (AD), (MST), and (USB) again mark the limits of identifiability.

Causality prospects. For causal discovery, recovering a structural equation model (SEM) on top of the identified (f, p_Z) is a further, nontrivial step: the analogue of the gap the Linear Non-Gaussian Acyclic Model (LiNGAM) [38] closed between linear ICA and linear SEM. For causal inference, in Intact-VAE [45], which is based on domain contrast, a central issue is to ensure that the *mechanisms of two treatment arms* are identified up to the *same* affine ambiguity—precisely the problem solved by our MID (Theorem 8).

Limitations. Results are at the population level (exact equality (1), the global optimum), as in iVAE and Kivva et al. [26]. Estimation is not claimed. The results assume equal latent and observation dimensions (a lower latent dimension, with or without noise, is not covered) and invertible branch rules. (SB)/(USB) genericity is conjectural under continuity. (AD) is sufficient, with necessity open. The causal reading is an analogy, identifying no causal graph, intervention or SEM. Proposition 12 has no rate and says nothing at margin zero.

Acknowledgements and disclosures

I thank my advisor, Prof. Shohei Shimizu, for organizing a seminar on this work, for his support since I joined the lab, and for the freedom to write this paper the way I wanted. I thank Thong Pham and Juraj Bodik for valuable discussions. I thank Robert A. Vandermeulen for bringing Kivva et al. [26] to my attention. I also thank the review committee for valuing purely theoretical work, and the program chairs for the theory-track reviewer guidelines, which let me argue the theory on its own terms, even in the rebuttal. This work is currently supported by JST CREST, Grant Number JPMJCR22D2, via my present lab; earlier stages were supported by IBS lab, Kyoto University. The author declares no competing interests.

References

[1] Kartik Ahuja, Jason Hartford, and Yoshua Bengio. Properties from mechanisms: An equivariance perspective on identifiable representation learning. In *International Conference on Learning Representations*, 2022.

[2] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37(6A):3099–3132, 2009.

[3] T. W. Anderson and Herman Rubin. Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 5:111–150, 1956.

[4] Jack Brady, Roland S. Zimmermann, Yash Sharma, Bernhard Schölkopf, Julius von Kügelgen, and Wieland Brendel. Provably learning object-centric representations. In *International Conference on Machine Learning*, 2023.

[5] Jack Brady, Julius von Kügelgen, Sébastien Lachapelle, Simon Buchholz, Thomas Kipf, and Wieland Brendel. Interaction asymmetry: A general principle for learning composable abstractions. In *International Conference on Learning Representations*, 2025. arXiv:2411.07784.

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

[7] Rizwan A. Choudrey. *Variational Methods for Bayesian Independent Component Analysis*. DPhil thesis, University of Oxford, 2002.

[8] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3): 287–314, 1994.

[9] Georges Darmois. Analyse générale des liaisons stochastiques. étude particulière de l'analyse factorielle linéaire. *Revue de l'Institut International de Statistique*, 21(1/2):2–8, 1953.

[10] Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, Matthew C. H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with Gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[11] Daniel Durstewitz. A state space approach for piecewise-linear recurrent neural networks for identifying computational dynamics from neural measurements. *PLoS Computational Biology*, 13(6):e1005542, 2017.

[12] Morris L. Eaton. *Group Invariance Applications in Statistics*, volume 1 of *Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and American Statistical Association, 1989.

[13] Fabian Falck, Haoting Zhang, Matthew Willetts, George Nicholson, Christopher Yau, and Chris C. Holmes. Multi-facet clustering variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 34, pages 8676–8690, 2021.

[14] Oren Freifeld, Søren Hauberg, Kayhan Batmanghelich, and John W. Fisher III. Transformations based on continuous piecewise-affine velocity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2496–2509, 2017.

[15] Narayan C. Giri. *Group Invariance in Statistical Inference*. World Scientific, 1996.

[16] Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete Rosetta Stone problem: Identifiability results for multi-view nonlinear ICA. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

[17] 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*, 2021.

[18] Christian Hennig. Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17(2):273–296, 2000.

[19] Irina Higgins, David Amos, David Pfau, Sébastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. In *arXiv preprint arXiv:1812.02230*, 2018.

[20] Leonid Hurwicz. Generalization of the concept of identification. In Tjalling C. Koopmans, editor, *Statistical Inference in Dynamic Economic Models*, number 10 in Cowles Commission Monograph, pages 245–257. Wiley, 1950.

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

[22] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

[23] Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[24] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1965–1972, 2017.

[25] Ilyes Khemakhem, Diederik P. Kingma, Ricardo P. Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[26] Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. In *Advances in Neural Information Processing Systems*, 2022.

[27] Joseph B. Kruskal. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2): 95–138, 1977.

[28] Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. In *Advances in Neural Information Processing Systems*, 2023.

[29] Scott W. Linderman, Matthew J. Johnson, Andrew C. Miller, Ryan P. Adams, David M. Blei, and Liam Paninski. Bayesian learning and inference in recurrent switching linear dynamical systems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[30] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019.

[31] Romain Lopez, Jan-Christian Huetter, Ehsan Hajiramezanali, Jonathan K. Pritchard, and Aviv Regev. Toward the identifiability of comparative deep generative models. In *Conference on Causal Learning and Reasoning (CLeaR)*, 2024.

[32] Stefan Matthes, Zhiwei Han, and Hao Shen. Mechanistic independence: A principle for identifiable disentangled representations. *arXiv preprint arXiv:2509.22196*, 2025.

[33] Guido F. Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, 2014.

[34] Simone Paoletti, Aleksandar Lj. Juloski, Giancarlo Ferrari-Trecate, and René Vidal. Identification of hybrid systems: A tutorial. *European Journal of Control*, 13(2–3):242–260, 2007.

[35] Geoffrey Roeder, Luke Metz, and Diederik P. Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, 2021.

[36] Daniel Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid registration using free-form deformations: Application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

[37] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[38] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006. URL `http://jmlr.org/papers/v7/shimizu06a.html`.

[39] V. P. Skitovich. On a property of the normal distribution. *Doklady Akademii Nauk SSSR*, 89: 217–219, 1953.

[40] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, 32(7):1153–1190, 2013.

[41] Chandler Squires and Pradeep Ravikumar. A unifying framework for unsupervised concept extraction. *arXiv preprint arXiv:2604.24936*, 2026.

[42] Feng Su and Yonglu Tian. Non-rigid registration for two-photon imaging using triangulation and piecewise affine transformation. *Neuroscience*, 491:86–99, 2022.

[43] Henry Teicher. Identifiability of finite mixtures. *Annals of Mathematical Statistics*, 34(4): 1265–1269, 1963.

[44] René Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, 2005.

[45] Pengzhou Abel Wu and Kenji Fukumizu. β-intact-VAE: Identifying and estimating causal effects under limited overlap. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=q7n2RngwOM`.

[46] Quanhan Xi and Benjamin Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

[47] Danru Xu, Sébastien Lachapelle, and Sara Magliacane. Identifiability of potentially degenerate gaussian mixture models with piecewise affine mixing, 2026. arXiv:2604.13218.

[48] Sidney J. Yakowitz and John D. Spragins. On the identifiability of finite mixtures. *Annals of Mathematical Statistics*, 39(1):209–214, 1968.

[49] Xiaojiang Yang, Yi Wang, Jiacheng Sun, Xing Zhang, Shifeng Zhang, Zhenguo Li, and Junchi Yan. Nonlinear ICA using volume-preserving transformations. In *International Conference on Learning Representations*, 2022.

[50] Yujia Zheng and Kun Zhang. Generalizing nonlinear ICA beyond structural sparsity. In *Advances in Neural Information Processing Systems*, 2023.

[51] Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. In *Advances in Neural Information Processing Systems*, volume 35, pages 16411–16422, 2022.

A Proofs of main results

Lemma 13 (Single-chamber PWA is globally affine)

Lemma 13. *Let $h\colon \mathbb{R}^d \to \mathbb{R}^d$ be a reduced PWA map with invertible affine branch rules. If h has exactly one chamber, then h has exactly one branch, hence $h(x) = Ax + b$ globally, up to the harmless measure-zero boundary convention for PWA maps.*

Proof. Write the branch rules as

$$h_i(z) = A_i z + b_i, \qquad A_i \in \mathrm{GL}(d),$$

and write

$$U_i := h_i(P_i)$$

for the image of branch i. Since each P_i is a finite union of full-dimensional polyhedra and h_i is an affine bijection, each U_i is again a finite union of full-dimensional polyhedra.

Let

$$\mathcal{B} := \bigcup_i \partial U_i.$$

The set $\mathcal{B}$ is contained in a finite union of lower-dimensional polyhedral pieces, hence has Lebesgue measure zero. On every connected component of

$$\mathbb{R}^d \setminus \mathcal{B},$$

membership in each U_i is constant, and therefore the active set

$$\mathcal{I}^h(x) = \{i : x \in U_i\}$$

is constant. Thus every such connected component is contained in a chamber.

By assumption there is exactly one chamber; call it C. Hence all points outside C are contained in $\mathcal{B}$. In particular,

$$\lambda_d(\mathbb{R}^d \setminus C) = 0,$$

so C has full Lebesgue measure.

We now show that every branch is active on C. Fix a branch i. Since P_i contains a full-dimensional polyhedral piece and h_i is invertible affine, U_i has nonempty interior. Therefore $U_i \setminus \mathcal{B}$ is nonempty. Pick

$$y \in U_i \setminus \mathcal{B}.$$

Then y lies in some chamber, and since there is only one chamber, $y \in C$. Also $y \in U_i$, so branch i is active at y. Since the active set is constant on C, branch i is active at every point of C. Equivalently,

$$C \subset U_i = h_i(P_i).$$

Because h_i is an affine bijection,

$$h_i^{-1}(C) \subset P_i.$$

Since C has full measure and affine bijections preserve null sets, $h_i^{-1}(C)$ also has full measure. Hence P_i has full measure.

Thus every branch domain P_i has full measure. If there were two distinct branches $i \neq j$, then $P_i \cap P_j$ would have full measure, contradicting the assumption that the branch domains form a partition up to measure-zero boundaries. Therefore there is exactly one branch.

With only one branch, the PWA map is given by a single invertible affine rule

$$h(x) = Ax + b$$

on the whole domain, up to the standard measure-zero boundary convention. Enlarging the unique branch domain by the remaining boundary points if necessary gives the stated global affine representative. □

Lemma 1 (Active-branch-count equality)

Proof. Fix a fine chamber C. On C, equation (2) gives an identity between two finite Gaussian mixtures. Each side is a finite sum of globally-defined Gaussian densities, hence a real-analytic function on $\mathbb{R}^d$. Since the two analytic functions agree on the nonempty open set C, they agree on all of $\mathbb{R}^d$.

Now integrate both sides over $\mathbb{R}^d$. Each Gaussian density integrates to 1, so every active branch contributes exactly one full mixture mass budget:

$$\text{LHS} = \sum_{i \in \mathcal{I}^f(C)} \sum_{k=1}^{K} w_k = \sum_{i \in \mathcal{I}^f(C)} 1 = |\mathcal{I}^f(C)|,$$

$$\text{RHS} = \sum_{j \in \mathcal{I}^g(C)} \sum_{\ell=1}^{K'} w'_\ell = \sum_{j \in \mathcal{I}^g(C)} 1 = |\mathcal{I}^g(C)|.$$

Therefore

$$|\mathcal{I}^f(C)| = |\mathcal{I}^g(C)|.$$

□

Lemma 2 (Global parameter bijection under (AD))

Proof. **Step 1: Chamber-wise matching.** Fix a fine chamber C. Equation (2) equates two finite Gaussian mixtures on the open set C. By Remark 4.1, this equality extends to all of $\mathbb{R}^d$.

Under (AD) on each side, the parameter families

$$\{\theta^{(f)}_{i,k}\}_{i \in \mathcal{I}^f(C),\, k} \qquad \text{and} \qquad \{\theta^{(g)}_{j,\ell}\}_{j \in \mathcal{I}^g(C),\, \ell}$$

are pairwise distinct within their respective sides. Hence the Gaussian-independence lemma applies to the difference of the two mixtures: every Gaussian parameter appearing on one side must also appear on the other, and matched terms must carry the same coefficient. Because all coefficients are strictly positive (by irredundancy of the GMMs), this gives a bijection

$$\Psi_C \colon \mathcal{I}^g(C) \times [K'] \;\longleftrightarrow\; \mathcal{I}^f(C) \times [K]$$

such that

$$\theta^{(g)}_{j,\ell} = \theta^{(f)}_{i,k}, \qquad w'_\ell = w_k,$$

whenever $\Psi_C(j, \ell) = (i, k)$.

Step 2: Consistency across chambers. We now show these chamber-wise bijections glue to a single global bijection.

Take any estimator pair $(j, \ell) \in I_g \times [K']$. Choose a fine chamber C with $j \in \mathcal{I}^g(C)$. Such a chamber exists because Q_j contains a full-dimensional polyhedron, and the invertible affine map g_j sends it to a set with nonempty interior; any interior point lies in some fine chamber with branch j active. Define

$$\Psi(j, \ell) := \Psi_C(j, \ell).$$

If C' is another fine chamber with $j \in \mathcal{I}^g(C')$, then $\Psi_C(j, \ell)$ and $\Psi_{C'}(j, \ell)$ are both truth-side pairs having parameter $\theta^{(g)}_{j,\ell}$. By (AD) on the truth side, there is *at most one* truth pair with that parameter. Hence

$$\Psi_C(j, \ell) = \Psi_{C'}(j, \ell),$$

so the definition of $\Psi(j, \ell)$ is independent of the chosen chamber.

The same uniqueness argument on the estimator side shows that the inverse matching is also well-defined. Therefore Ψ is a global bijection

$$\Psi \colon I_g \times [K'] \;\longleftrightarrow\; I_f \times [K]$$

with the stated parameter and weight matching property, and whose restriction to each fine chamber is exactly Ψ_C. □

Lemma 14 (Component-count equality)

Lemma 14 (Component-count equality)**.** $K' = K$.

Proof. Choose a fine chamber C with nonempty active set; such a chamber exists because each branch image has nonempty interior. On C, the chamber-wise bijection Ψ_C matches

$$|\mathcal{I}^f(C)| \cdot K$$

truth-side pairs to

$$|\mathcal{I}^g(C)| \cdot K'$$

estimator-side pairs. Hence

$$|\mathcal{I}^f(C)| \cdot K = |\mathcal{I}^g(C)| \cdot K'.$$

By Lemma 1,

$$|\mathcal{I}^f(C)| = |\mathcal{I}^g(C)| \geq 1.$$

Dividing by this common positive number gives

$$K = K'.$$

□

Lemma 15 (Boundary branch-count equality)

Lemma 15 (Boundary branch-count equality)**.** *At every pair of adjacent fine chambers,* $D_f(C, C') = D_g(C, C')$ *and* $E_f(C, C') = E_g(C, C')$.

Proof. Let C, C' be adjacent fine chambers. Define the leaving blocks

$$L_g := \big(\mathcal{I}^g(C) \setminus \mathcal{I}^g(C')\big) \times [K], \qquad L_f := \big(\mathcal{I}^f(C) \setminus \mathcal{I}^f(C')\big) \times [K].$$

We claim that Ψ restricts to a bijection $L_g \longleftrightarrow L_f$. Take any $(j, \ell) \in L_g$. Then j is active on C but not on C', so the Gaussian with parameter $\theta^{(g)}_{j,\ell}$ appears in the chamber identity for C but not in the chamber identity for C'. Let

$$\Psi(j, \ell) = (i, k).$$

Then $\theta^{(f)}_{i,k} = \theta^{(g)}_{j,\ell}$. Since (j, ℓ) is active on C, the pair (i, k) must be active on C as well, by the chamber-wise matching on C.

If (i, k) were also active on C', then the same Gaussian parameter would appear on the truth side for C'. But on C', chamber-wise matching would then force the corresponding estimator pair with that parameter to be active there too. By global uniqueness under (AD), that estimator pair must be (j, ℓ), contradicting $(j, \ell) \notin \mathcal{I}^g(C') \times [K]$. Hence $(i, k) \in L_f$.

Thus $\Psi(L_g) \subseteq L_f$. Applying the same argument with the roles of f and g reversed gives the reverse inclusion, so Ψ indeed restricts to a bijection

$$L_g \longleftrightarrow L_f.$$

Now count elements:

$$|L_g| = D_g \cdot K, \qquad |L_f| = D_f \cdot K.$$

Since the two sets are in bijection,

$$D_g \cdot K = D_f \cdot K.$$

Because $K \geq 1$, we obtain $D_g = D_f$. The entering-count identity $E_g = E_f$ is proved in exactly the same way, after swapping C and C'. □

Lemma 16 (Chamber coincidence)

Lemma 16 (Chamber coincidence)**.** *Under double (AD), let C, C' be adjacent fine chambers with common boundary F. Then $\mathcal{I}^f(\cdot)$ changes across F if and only if $\mathcal{I}^g(\cdot)$ does. Hence f-chambers, g-chambers, and fine chambers all coincide and have the same boundaries.*

Proof. The boundary F is, by definition, a locus where at least one of the active sets changes. Equivalently, across F either $D_f + E_f > 0$ or $D_g + E_g > 0$ (or both). Lemma 15 gives

$$D_f = D_g, \qquad E_f = E_g,$$

hence

$$D_f + E_f > 0 \iff D_g + E_g > 0.$$

So a fine-chamber boundary is a boundary for f if and only if it is one for g.

Now let B be any boundary of an f-chamber. By definition, crossing B changes $\mathcal{I}^f(\cdot)$, so B is also a fine-chamber boundary. By the equivalence above, crossing B also changes $\mathcal{I}^g(\cdot)$, so B is a boundary of a g-chamber. The same argument with the roles reversed shows that every boundary of a g-chamber is also one of an f-chamber.

Therefore the f-chambers, the g-chambers, and the fine chambers all coincide. Equalities of branch domains and of unions of chambers are up to Lebesgue-null sets (Definitions 3.2 and 3.3). □

Theorem 4 (From PWA to affine: the symmetry-rigidity)

Proof. Take the truth map to be $f = \mathrm{Id}$. Because Id has a single chamber, (Id, Z) trivially satisfies (AD), so double (AD) holds.

The identity map has exactly one chamber, namely all of $\mathbb{R}^d$, and its active set on that chamber has size 1. By Lemma 16, g also has a single chamber. By Lemma 1, the active set of g on that chamber also has size 1.

Since every branch of a reduced PWA map is active on some open subset of its image, a map with only one chamber and active-set size 1 must in fact have only one branch in total (Lemma 13). A one-branch PWA map is globally affine, so

$$g(x) = Ax + b$$

for some $A \in \mathrm{GL}(d)$ and $b \in \mathbb{R}^d$.

The "in particular" clause (no strict PWA self-maps) follows by taking $f = \mathrm{Id}$ and $Z \sim p$, so that the observable law $g(Z) \sim Z$ is itself a GMM. □

Theorem 5 (Law identifiability)

Proof. Lemma 14 gives

$$K' = K =: K.$$

Case 1: f has a single chamber. Then f is globally affine. Write

$$f(x) = Ax + b, \qquad A \in \mathrm{GL}(d).$$

Hence $f(Z)$ is itself a GMM. Since

$$g(Z') \sim f(Z)$$

and (g, Z') satisfies (AD), Theorem 4 implies that g is also globally affine, say

$$g(x) = Bx + c.$$

Define

$$\alpha := f^{-1} \circ g.$$

Then α is an affine bijection, and

$$\alpha(Z') = (f^{-1} \circ g)(Z') = f^{-1}\big(g(Z')\big) \sim f^{-1}\big(f(Z)\big) = Z.$$

So LID holds in this case.

Case 2: f has more than one chamber. By Lemma 16, the f-chambers are exactly the fine chambers. Thus an (SB) boundary for f is a fine-chamber boundary.

By (SB), there exist adjacent chambers C, C' and a unique toggling truth branch. Swapping C and C' if necessary, we may assume

$$\mathcal{I}^f(C) \setminus \mathcal{I}^f(C') = \{i_\star\},$$

so $D_f = 1$. Lemma 15 then gives $D_g = 1$, so there is a unique estimator branch $j_\star$ leaving across the same boundary.

The global bijection Ψ restricts to a bijection between the leaving blocks

$$\{j_\star\} \times [K] \longleftrightarrow \{i_\star\} \times [K].$$

Since both sets have size K, this restricted bijection is equivalent to a permutation $\sigma\colon [K] \to [K]$ such that, for every $\ell \in [K]$,

$$\theta^{(g)}_{j_\star,\ell} = \theta^{(f)}_{i_\star,\sigma(\ell)}, \qquad w'_\ell = w_{\sigma(\ell)}.$$

Unpacking the equality of Gaussian parameters, we get

$$B_{j_\star}\mu'_\ell + c_{j_\star} = A_{i_\star}\mu_{\sigma(\ell)} + b_{i_\star},$$
$$B_{j_\star}\Sigma'_\ell B_{j_\star}^\top = A_{i_\star}\Sigma_{\sigma(\ell)}A_{i_\star}^\top.$$

Because both branch matrices are invertible, the map

$$\alpha(z) := A_{i_\star}^{-1}\big(B_{j_\star}z + c_{j_\star} - b_{i_\star}\big)$$

is an affine bijection.

The two displayed identities say exactly that α sends the estimator component $\mathcal{N}(\mu'_\ell, \Sigma'_\ell)$ to the truth component $\mathcal{N}(\mu_{\sigma(\ell)}, \Sigma_{\sigma(\ell)})$, and the matched weights are equal. Therefore

$$\alpha(Z') = \sum_{\ell=1}^{K} w'_\ell \mathcal{N}(\mu_{\sigma(\ell)}, \Sigma_{\sigma(\ell)}) \sim Z.$$

This proves LID. □

Proposition 6 (((SB)) and ((USB)) inheritance)

Proof. **((SB)) case.** If f has a single chamber, Lemma 16 gives the same for g, and clause (a) applies. If f is strict PWA and satisfies (SB), there exist adjacent chambers with a unique truth toggling branch (say leaving); Lemma 15 gives $D_g = 1$, so g also has a simple boundary. Hence g satisfies (SB).

((USB)) case. If f has a single chamber, then by Lemma 16 so does g, and clause (a) of (USB) holds for g automatically.

Assume now that f is strict PWA. Fix any truth branch $i \in I_f$. Because f satisfies (USB), there exist adjacent chambers where i is the unique leaving branch or the unique entering branch. Swapping the chambers if needed, assume i is the unique leaving branch. Then $D_f = 1$, and Lemma 15 gives $D_g = 1$. So there is a unique estimator branch, call it $j(i)$, that toggles in the same direction across that boundary.

We claim that the map $i \mapsto j(i)$ is injective. Suppose, toward contradiction, that

$$j(i_1) = j(i_2) = j_\star$$

for two distinct truth branches $i_1 \neq i_2$. At the boundary exposing i_1, the leaving block argument shows that Ψ restricts to a bijection

$$\{j_\star\} \times [K] \longleftrightarrow \{i_1\} \times [K].$$

At the boundary exposing i_2, the same global Ψ restricts to

$$\{j_\star\} \times [K] \longleftrightarrow \{i_2\} \times [K].$$

But for any fixed ℓ, the pair $\Psi(j_\star, \ell)$ is single-valued. Its first coordinate therefore cannot be both i_1 and i_2. Contradiction.

Thus $i \mapsto j(i)$ is injective. Since $|I_f| = |I_g|$, an injection between these finite sets is automatically a bijection. So every estimator branch appears as some $j(i)$ and is exposed at its own simple boundary. Therefore g satisfies (USB). □

Proposition 17 ((MST) conjugacy)

Proposition 17 ((MST) conjugacy)**.** *For any affine bijection α with $\alpha(Z') \sim Z$, $G_{\mathrm{mix}}(p_{Z'}) = \alpha^{-1} G_{\mathrm{mix}}(p_Z)\, \alpha$. In particular, p_Z satisfies (MST) if and only if $p_{Z'}$ does.*

Proof. Let $T \in \mathrm{Aff}(\mathbb{R}^d)$. Then

$$T_\sharp p_{Z'} = p_{Z'}$$

if and only if

$$\alpha_\sharp T_\sharp p_{Z'} = \alpha_\sharp p_{Z'}.$$

Using $\alpha_\sharp p_{Z'} = p_Z$, this is equivalent to

$$(\alpha T \alpha^{-1})_\sharp p_Z = p_Z.$$

So

$$T \in G_{\mathrm{mix}}(p_{Z'}) \iff \alpha T \alpha^{-1} \in G_{\mathrm{mix}}(p_Z).$$

Hence

$$G_{\mathrm{mix}}(p_{Z'}) = \alpha^{-1}\, G_{\mathrm{mix}}(p_Z)\, \alpha.$$

In particular, one symmetry group is trivial if and only if the other is. □

Lemma 7 (MID with shared latent)

Proof. From Section 4.1 we already know:

$$K = K', \qquad |I_f| = |I_g|,$$

the leaving and entering counts agree at every boundary (Lemma 15), and the chamber decompositions coincide (Lemma 16).

Step 1: Match each truth branch to one estimator branch. Fix any truth branch $i \in I_f$.

If f has a single chamber, then by chamber coincidence so does g. Hence both maps are globally affine. Then

$$S := f^{-1} \circ g$$

is affine and satisfies

$$S_\sharp\, p = {f^{-1}}_\sharp\, g_\sharp\, p \sim {f^{-1}}_\sharp\, f_\sharp\, p = p.$$

So $S \in G_{\mathrm{mix}}(p)$, and (MST) gives $S = \mathrm{Id}$, hence $f = g$. Thus the single-chamber case is done.

Assume from now on that f is strict PWA. By (USB), there exist adjacent chambers C, C' such that i is the unique leaving or entering truth branch. Swapping the chambers if needed, we may assume i is the unique leaving branch, so $D_f = 1$. Lemma 15 gives $D_g = 1$; let $j(i) \in I_g$ be the unique estimator branch that leaves across the same boundary.

The restricted bijection between leaving blocks

$$\Psi \colon \{j(i)\} \times [K] \longleftrightarrow \{i\} \times [K]$$

induces a permutation $\sigma_i \colon [K] \to [K]$ such that

$$\theta^{(g)}_{j(i),\ell} = \theta^{(f)}_{i,\sigma_i(\ell)}, \qquad w_\ell = w_{\sigma_i(\ell)} \tag{4}$$

for all $\ell \in [K]$.

Step 2: The branch difference is a mixture symmetry. Define

$$S_i := f_i^{-1} \circ g_{j(i)}.$$

Because the latent law is the same on both sides, the parameter identity (4) says that S_i sends the ℓ-th Gaussian component to the $\sigma_i(\ell)$-th one:

$$S_{i\sharp}\, \mathcal{N}_\ell = \mathcal{N}_{\sigma_i(\ell)}.$$

The weight identity in (4) shows that this permutation preserves mixture weights. Therefore

$$S_{i\sharp}\, p = \sum_{\ell=1}^{K} w_\ell \, \mathcal{N}_{\sigma_i(\ell)} = \sum_{m=1}^{K} w_m \, \mathcal{N}_m = p.$$

Hence $S_i \in G_{\mathrm{mix}}(p)$.

Step 3: Kill the ambiguity with (MST). By (MST),

$$G_{\mathrm{mix}}(p) = \{\mathrm{Id}\}.$$

So $S_i = \mathrm{Id}$, which means

$$f_i = g_{j(i)}$$

for every truth branch i.

Step 4: Convert this into equality of reduced PWA maps. The correspondence $i \mapsto j(i)$ is well-defined. It is injective by exactly the same argument as in Proposition 6: one estimator branch cannot be the unique toggling partner of two different truth branches without making Ψ multivalued on some (j, ℓ). Since $|I_f| = |I_g|$, this injection is a bijection; write it as

$$\pi^{-1} \colon I_f \to I_g, \qquad \pi(j(i)) = i.$$

It remains to compare branch domains. Let $i \in I_f$. For any fine chamber C, active-set agreement together with the bijection π gives

$$i \in \mathcal{I}^f(C) \iff \pi^{-1}(i) \in \mathcal{I}^g(C).$$

By definition of active set, this means

$$C \subset f_i(P_i) \iff C \subset g_{\pi^{-1}(i)}(Q_{\pi^{-1}(i)}).$$

Branch domains may be open, closed or mixed, and chambers are open (Definition 3.4). Equalities of branch domains and of unions of chambers are up to Lebesgue-null sets (Definitions 3.2 and 3.3). Taking the union over all such fine chambers recovers the interiors of the two branch images up to a Lebesgue-null set:

$$f_i(P_i)^\circ = g_{\pi^{-1}(i)}(Q_{\pi^{-1}(i)})^\circ.$$

But Step 3 gave $f_i = g_{\pi^{-1}(i)}$, and this common map is an invertible affine bijection. Applying its inverse, which preserves interiors, yields

$$P_i^\circ = Q_{\pi^{-1}(i)}^\circ.$$

Finally, each branch domain is a finite union of full-dimensional polyhedra, hence differs from its interior by a Lebesgue-null set. Therefore

$$P_i = Q_{\pi^{-1}(i)}$$

up to measure-zero boundaries. By Definition 3.3, $f = g$. □

Theorem 8 (Map identifiability)

Proof. **Phase I: Obtain the latent affine link.** Since (USB) implies (SB), Theorem 5 yields an affine bijection α and a permutation σ of $[K]$ such that

$$\alpha(Z') \sim Z$$

and, componentwise,

$$\alpha_\sharp \mathcal{N}(\mu'_\ell, \Sigma'_\ell) = \mathcal{N}(\mu_{\sigma(\ell)}, \Sigma_{\sigma(\ell)}), \qquad w'_\ell = w_{\sigma(\ell)}. \tag{5}$$

Phase II: Reduce to a shared latent law. Define

$$h := g \circ \alpha^{-1}.$$

Since $\alpha(Z') \sim Z$, equivalently $\alpha^{-1}(Z) \sim Z'$, we get

$$h(Z) = (g \circ \alpha^{-1})(Z) = g\big(\alpha^{-1}(Z)\big) \sim g(Z') \sim f(Z).$$

So f and h push forward the *same* latent GMM Z to the same observable law.

We next check that (h, Z) still satisfies (AD). From (5),

$$(\alpha^{-1})_\sharp \mathcal{N}(\mu_k, \Sigma_k) = \mathcal{N}(\mu'_{\sigma^{-1}(k)}, \Sigma'_{\sigma^{-1}(k)}).$$

Therefore, for branch j of h and component k of Z, the pushforward parameters satisfy

$$\theta^{(h)}_{j,k} = \theta^{(g)}_{j,\sigma^{-1}(k)}. \tag{6}$$

Now suppose $\theta^{(h)}_{j,k} = \theta^{(h)}_{m,n}$. Using (6),

$$\theta^{(g)}_{j,\sigma^{-1}(k)} = \theta^{(g)}_{m,\sigma^{-1}(n)}.$$

Since (g, Z') satisfies (AD), we get

$$(j, \sigma^{-1}(k)) = (m, \sigma^{-1}(n)),$$

hence $(j, k) = (m, n)$. So (h, Z) satisfies (AD).

Also, composing a reduced PWA map with a global affine bijection preserves the reduced PWA structure: branch domains remain finite unions of full-dimensional polyhedra, invertibility of branch rules is unchanged, and distinct affine rules remain distinct after composing on the right with the same affine bijection. Thus (h, Z) is again a reduced PWA pair.

Phase III: Apply the shared-latent MID lemma. We can now apply Lemma 7 to the pair (f, h) with common latent law Z. The truth side (f, Z) still satisfies (USB) and (MST), and the estimator side (h, Z) satisfies (AD). Therefore

$$h = f.$$

Recalling $h = g \circ \alpha^{-1}$, we obtain

$$f = g \circ \alpha^{-1}.$$

□

Proposition 9 (Posterior identifiability)

Proof. We argue at the level of regular conditional distributions (posterior kernels), not Lebesgue densities: for a deterministic map the conditional law of the latent given X is typically supported on a finite fibre rather than on a full-dimensional set.

The subtlety is that LID/MID give only equality *in law*, $p_{Z'} = (\alpha^{-1})_\sharp\, p_Z$, not a coupling of the estimator latent Z' with the truth latent Z on a common space. We therefore build a *surrogate* estimator latent from the truth: on the truth space set

$$\widetilde{Z'} := \alpha^{-1}(Z), \qquad X := f(Z).$$

By the latent-law identity, $\widetilde{Z'} \sim Z'$, and by Theorem 8 ($f = g \circ \alpha^{-1}$, up to the null-boundary convention),

$$X = f(Z) = g\big(\alpha^{-1}(Z)\big) = g(\widetilde{Z'}) \qquad \text{almost surely.}$$

Thus $(\widetilde{Z'}, X)$ is generated by drawing $\widetilde{Z'}$ from the law of Z' and outputting $X = g(\widetilde{Z'})$—which is *exactly* how the estimator model generates its latent–observation pair $(Z', g(Z'))$. Hence the two pairs share the same joint law, so the estimator posterior $P_{\text{est}}(\cdot \mid x)$ (the conditional law of Z' given its observation) equals the conditional law of $\widetilde{Z'}$ given $X = x$, for P_X-almost every x.

It remains to identify the latter. Since $\widetilde{Z'} = \alpha^{-1}(Z)$ is a deterministic function of Z, for every Borel set $B \subseteq \mathbb{R}^d$ we have $\mathbf{1}_{\{\widetilde{Z'} \in B\}} = \mathbf{1}_{\{Z \in \alpha(B)\}}$ almost surely. Taking conditional expectations given X (with measurable versions of the regular conditional laws),

$$\underbrace{P(\widetilde{Z'} \in B \mid X = x)}_{= P_{\text{est}}(Z' \in B \mid x)} = \underbrace{P(Z \in \alpha(B) \mid X = x)}_{= P_{\text{true}}(Z \in \alpha(B) \mid x)}$$

for P_X-almost every x, which is exactly the conclusion of Proposition 9. □

Remark A.1. The only measure-theoretic subtlety is that posteriors are conditional probability kernels, hence defined only up to P_X-null sets. The proof above avoids formal disintegration language, but the underlying reason it works is simple: once two latent–observation pairs have the same joint distribution, they have the same posterior kernels, up to P_X-almost-everywhere equality.

Proposition 10 (Pointwise identifiability)

Proof. Injectivity of f implies that for every x in its image, the fiber $f^{-1}(x)$ is a singleton $\{z\}$. Since

$$f = g \circ \alpha^{-1} \qquad \Longleftrightarrow \qquad g = f \circ \alpha,$$

and α is a bijection, the map g is also injective. Hence

$$g^{-1}(x) = \{\alpha^{-1}(z)\}.$$

So if $\hat{z} := g^{-1}(x)$, then

$$\hat{z} = \alpha^{-1}(z),$$

equivalently

$$z = \alpha(\hat{z}).$$

□

Corollary 11 (PWA-ICA identifiability)

See Appendix J for the full proof.

B Terminology

Table 3: Compact terminology mapping (intuitive ↔ formal). Used consistently throughout the paper.

Intuitive term	Formal object	Notes
law	probability distribution, e.g. p_Z	*latent law* for p_Z; pushforward law for $f_\sharp\, p_Z$
map	decoder/generator f (reduced PWA throughout)	also called mixing function
domain	mixture component of latent GMM	not function domain; related to domain adaptation
mechanism	branch (affine rule) of reduced PWA map	each domain is acted on by every mechanism whose region it intersects
boundary	branch / chamber boundary	locus where the active branch set changes
contrast	algebraic genericity / non-symmetry	instances: (MST) (domain), (SB)/(USB) (mechanism), (AD) (interaction)

C Comparison of Related Work

D Linear independence of Gaussian densities

Lemma 18 (Linear independence of Gaussian densities [43])**.** *Let $\{\theta_r\}_{r=1}^M$ be pairwise distinct parameters $\theta_r = (m_r, S_r)$ with S_r positive definite. If $\sum_{r=1}^M \alpha_r\, \varphi_{\theta_r}(x) = 0$ for all $x \in \mathbb{R}^d$, then $\alpha_r = 0$ for all r.*

Proof sketch. Take Fourier transforms. The Fourier transform of a non-degenerate Gaussian is again a non-vanishing Gaussian; distinct mean–covariance pairs yield distinct exponential-quadratic forms. These are linearly independent as functions, so all coefficients vanish. A detailed proof can be found in classical GMM identifiability references. □

Table 4: Comparison of related work. Legend: ✓ allowed/supported; × ruled out / required not to happen; ∼ partial / under restrictions; n/a not the object of study. Rows ordered most-recent to oldest; ours on top. Types follow Matthes et al. [32]: **D**, the two factors' Jacobian columns occupy disjoint observation coordinates; **M**, supports may overlap but neither contains the other; **S**, the factor-aligned basis is the strict ℓ_0-sparsest; $\mathbf{H}_n$, the n-th mixed-derivative block between the factors vanishes; **O**, Jacobian columns inner-product orthogonal. D is the strongest. Our own row carries none of these tags, so no result changes. Kivva et al. [26] and Xu et al. [47] also have multiple PWA branches; the branch-structure column records whether that structure is used to identify, not whether it exists. On the domain-contrast axis, (MST) strictly weakens iVAE's observed labels plus assumption (iv): it can hold with $d+1$ components and no labels, where (iv), for Gaussian latents, needs $K > 2d$ (App. U).

Paper	**Model class** (f, Z)	**Constraint location**	**Proof flavour**	**Branch structure used as an identification resource**	**Discont.** f	**Non-inj.** f	**Non-indep. latent**	**Unsup. domains**
Ours	reduced PWA + irredundant GMM	Map ((USB)); Law ((MST)); Interaction ((AD))	algebraic + analytic continuation on chambers; symmetry-group collapse	✓ (branches of one f)	✓	✓ (except pointwise)	✓ (affine; ICA form under (DD))	✓ (mixture labels latent)
Xu et al. [47]	inj. continuous PWA + pdGMM	Map (inj. cont. PWA); Law (pdGMM genericity, shared basis); encoder sparsity	analytic + sparsity (extends Kivva past degeneracy)	×	×	×	∼ (low-rank / dependent latents via degeneracy)	✓
Matthes et al. [32]	generic nonlinear, possibly non-invertible	Interaction (support / sparsity / higher-order on f); law free	algebraic / graph-theoretic taxonomy unifying Type O/D/S/H	×	× (smooth)	∼ (local inj.)	✓	✓
Brady et al. [5] (Type H)	C^{n+1} generator + block latents	Interaction (block-diag. on $(n+1)$-th derivatives)	differential (higher-order Jacobian; autoencoder regulariser)	×	×	× (smooth autoencoder)	✓ (within-block dep.)	✓
Lopez et al. [31]	PWA inj. a.e. + structured GMM	Law (block structure across two conditions)	reduces to Kivva Thm. D.4; block-wise affine ID	× (two conditions sharing background)	×	× (∼ a.e. inj.)	∼ (within-block dep.; indep. across split)	× (observed condition)
Lachapelle et al. [28] (Type H)	C^2-diffeo additive decoder + block latents	Map (additivity + sufficient nonlinearity)	differential (block-wise)	×	×	× (C^2 diffeo)	✓ ("almost arbitrary support")	✓
Brady et al. [4] (Type D/S)	diffeo compositional decoder + slot latents	Interaction (compositionality + irreducibility on J_f)	differential (Jacobian sparsity)	× (slot blocks)	×	×	✓	✓
Zheng and Zhang [50] (Type M)	smooth, possibly non-bijective + grouped latents	Map (structural / partial sparsity); Law (partial dep.)	differential + sparsity	×	×	× (undercomplete, still inj.)	✓ (explicit)	✓
Xi and Bloem-Reddy [46]	injective f + unspecified prior family	Taxonomic: $\mathcal{A}(\mathcal{F}) \cap \mathcal{A}(\mathcal{P}_z)$	algebraic taxonomy; measure-theoretic intersection	×	n/a	× (Asn. 2)	n/a	× (multi-env. needs labels)
Buchholz et al. [6]	conformal / smooth f + indep. latents	Map (function-class restriction)	differential (function-class restriction)	×	×	×	×	✓
Yang et al. [49]	volume-preserving f + latents cond. indep. given a class	Map (volume preservation); Law (two overlapping classes)	differential	×	×	× (invertible flow)	× (cond. indep. given class)	× (class observed)
Kivva et al. [26]	continuous PWA + non-deg. GMM	Map (continuity + inj. on open f-image set / global inj.); Law (GMM)	analytic (continuation, GMM matching)	×	∼ (cont. for MID; LID allows discont.)	∼ (open-set inj. for LID; global inj. for MID)	∼ (no separation of MID from ICA-amb.)	✓
Ahuja et al. [1]	diffeo decoder + *known* dynamics m_t	Map (diffeo); known m_t / hypothesis class	group / equivariance (commutation)	× (iterated single g)	×	×	✓	× (time observed; mechanism known)
Roeder et al. [35]	broad family of discriminative softmax-form models	distributional matching of conditionals + diversity	algebraic (linear ID in function space)	×	n/a	✓ (no inj.)	✓	n/a
Gresele et al. [17] (IMA, Type O)	smooth invertible f + indep. latents	Interaction (Jacobian columns $\perp$)	differential (Jacobian-based; not a full ID theorem)	×	×	×	×	✓
Khemakhem et al. [25] (iVAE)	smooth invertible f + EF prior cond. on aux	Map (smooth/inv.); Law (EF + aux variability)	differential (Jacobian + sufficient stats)	×	×	×	× (cond. indep. given aux)	× (aux observed)
Gresele et al. [16] (Multi-View NICA)	smooth invertible views + indep. noise	Law (indep. + SDV on cond. log-density); Map (smooth/inv.)	differential (log-density factorisation, contrastive)	∼ (multi-view)	×	×	×	∼ (no env. labels but paired views)
Higgins et al. [19]	encoder $f: W \to Z$; group acts on W	direct-product decomp. *given*; latent law plays no role	definitional, group / representation theory	n/a	n/a	hand-waved	∼ (factor-as-subgroup)	n/a
Hyvärinen and Morioka [21] (TCL)	smooth invertible f + non-stationary indep. latents	Map (smooth/inv.); Law (segment variability)	differential (Jacobian / log-density, contrastive)	×	×	×	× (indep. required)	× (segment labels)

E Global vs. chamber-wise (AD)

Remark E.1 (Global vs. chamber-wise). Chamber-wise (AD)—injectivity of the parameter map restricted to the active pairs on each individual chamber—is genuinely weaker than the global condition stated above. The glueing step in Lemma 2 requires the *global* version: uniqueness of the match (i, k) for a given parameter value across *all* branches, not just those active on a single chamber. A counterexample showing that injectivity of the PWA map does not imply global (AD) (even though it implies chamber-wise (AD)) appears below.

Chamber-wise (AD) requires only that the parameter map $(i, k) \mapsto \theta^{(f)}_{i,k}$ be injective when restricted to the active pairs $\mathcal{I}^f(C) \times [K]$ on each individual chamber C. Global (AD) requires injectivity over *all* branch-component pairs. The following example shows these are genuinely different.

Proposition 19. *There exists an injective PWA map f and a GMM Z such that (f, Z) satisfies chamber-wise (AD) but not global (AD).*

Proof. Take $d = 1$, two branches, two mixture components. Define

$$f_1(z) = z \text{ on } P_1 = (-\infty, 0), \qquad f_2(z) = z + 10 \text{ on } P_2 = (0, \infty).$$

This f is injective and reduced (distinct affine rules). Let

$$Z \sim w_1 \mathcal{N}(\mu_1, \sigma^2) + w_2 \mathcal{N}(\mu_2, \sigma^2),$$

with $\mu_2 = \mu_1 - 10$. Then

$$\theta^{(f)}_{1,1} = (\mu_1, \sigma^2) = (\mu_2 + 10, \sigma^2) = \theta^{(f)}_{2,2},$$

so global (AD) fails.

On the other hand, since f is injective, each nonempty chamber has exactly one active branch. Hence the active pairs on any chamber form $\{i\} \times [K]$ for a single i, and the restriction of the parameter map to that set is injective by invertibility of the branch rule and irredundancy of the GMM. So chamber-wise (AD) holds. □

Where global (AD) enters. The glueing step in Lemma 2 uses the statement: for fixed (j, ℓ), there is a *unique* (i, k) with $\theta^{(f)}_{i,k} = \theta^{(g)}_{j,\ell}$ across all branches, not just those active on a given chamber. Chamber-wise (AD) alone allows a "hidden duplicate parameter in a different branch" phenomenon: on chamber C with active set $\{i\}$, the match gives $\Psi_C(j, \ell) = (i, k)$; but if $\theta^{(f)}_{i,k} = \theta^{(f)}_{i',k'}$ for some $i' \neq i$ active on a different chamber C', then $\Psi_{C'}(j, \ell)$ could yield (i', k'), breaking the global consistency of Ψ. This inconsistency propagates: the boundary-counting argument in Lemma 15 and the "leaving blocks" tracking in subsequent results all rely on the global bijection being well-defined.

F (AD) hierarchy, dimension counting, and information view

Let $D := d(d-1)/2$ denote the dimension of the orthogonal symmetries of a covariance matrix (the maximal dimension of the component symmetry group; see Appendix G.1).

(AD) decomposes naturally into two sub-conditions, which together with the orbit-based conditions (OD) and (TOD) form a hierarchy of increasingly strong genericity requirements on the map–latent pair.

CROSSCOMPAD (cross-component). A violation $\theta_{j,\ell} = \theta_{m,n}$ with $\ell \neq n$ forces $h_m^{-1} \circ h_j$ to *transport* $\mathcal{N}_\ell$ onto $\mathcal{N}_n$. The corresponding bad set is a coset of $O(\Sigma_\ell)$ and has dimension exactly D.

SAMECOMPAD (same-component). A violation $\theta_{j,\ell} = \theta_{m,\ell}$ with $j \neq m$ means the difference map $h_m^{-1} \circ h_j$ preserves $\mathcal{N}_\ell$, i.e., it is a *symmetry* of that component. The "bad" set of such affine maps has dimension exactly D.

(OD) (Cross-component orbit disjointness). For distinct branches $j \neq m$ and distinct components $\ell \neq n$, $[h_j]_\ell \cap [h_m]_n = \varnothing$. A violation means the difference map $h_m^{-1} \circ h_j$ lies in the product $\mathrm{Sym}_n \cdot \mathrm{Sym}_\ell^{-1}$, a set of dimension at most $2D$ inside $\mathrm{Aff}(\mathbb{R}^d)$ (dimension $d^2 + d$). The "bad set" has codimension at least $2d$, hence measure zero.

(TOD) (Transporter-orbit disjointness). For all distinct branches $i \neq i'$ and all component quadruples (k, ℓ, k', ℓ'), $\big(h_i \circ \mathrm{Aff}(k \to \ell)\big) \cap \big(h_{i'} \circ \mathrm{Aff}(k' \to \ell')\big) = \varnothing$. Setting $k = \ell$ and $k' = \ell'$ reduces $\mathrm{Aff}(k \to k) = \mathrm{Sym}_k$, so (TOD) implies (OD) (further restricted to $k \neq k'$) and SAMECOMPAD (further restricted to $k = k'$). (TOD) is thus strictly stronger than both.

Dimension counting. Both CROSSCOMPAD and SAMECOMPAD bad sets sit inside $\mathrm{Aff}(\mathbb{R}^d)$ (dimension $d^2 + d$) with positive codimension, hence measure zero. While their dimensions match, CROSSCOMPAD and SAMECOMPAD differ in what the difference map must "know": cross-component collisions require the difference between two branches to encode a relationship between *two* independent Gaussian components, while same-component collisions require it to encode the *self-symmetry* of a single component.

Algorithmic-information viewpoint. For (OD), the dimension doubles because a joint orbit intersection involves two independent degrees of freedom, one for each component's symmetry group. Crucially, the difference map need *not* fix either component individually; it only needs to realize the joint constraint coming from both orbits. This makes (OD) logically independent of SAMECOMPAD. For (TOD), a violation requires the difference map $h_{i'}^{-1} \circ h_i$ to lie in $\mathrm{Aff}(k' \to \ell')^{-1} \circ \mathrm{Aff}(k \to \ell)$, a set of dimension at most $2D$ (two independent orthogonal degrees of freedom, one from each covariance). The full affine group has dimension $d^2 + d$, so the bad set has codimension at least $2d$—the same bound as for (OD).

G Component symmetries and alternative conditions

G.1 Component symmetry groups and orbits

For each component ℓ of a GMM, the *component symmetry group* is

$$\mathrm{Sym}_\ell := \big\{T \in \mathrm{Aff}(\mathbb{R}^d) : \; T_\sharp \mathcal{N}_\ell = \mathcal{N}_\ell\big\}. \tag{7}$$

Concretely, $T(x) = Ax + b \in \mathrm{Sym}_\ell$ if and only if $A\Sigma_\ell A^\top = \Sigma_\ell$ and $b = (I - A)\mu_\ell$. Thus Sym_ℓ is isomorphic to $O(\Sigma_\ell)$, and has dimension $D := d(d-1)/2$.

The *orbit* of branch j under component ℓ is $[h_j]_\ell := \big\{h_j \circ T : \; T \in \mathrm{Sym}_\ell\big\}$.

Two branches have matching pushforward parameters on component ℓ if and only if they belong to the same orbit:

Proposition 20 (SAMECOMPAD $\iff$ orbit disjointness)**.** (h, Y) *satisfies* SAMECOMPAD *if and only if* $[h_j]_\ell \cap [h_m]_\ell = \varnothing$ *for all distinct branches* $j \neq m$ *and any component* ℓ*.*

Proof. The pushforward parameters match, $\theta_{j,\ell} = \theta_{m,\ell}$, iff $h_m^{-1} \circ h_j \in \mathrm{Sym}_\ell$. Rearranging: $h_j \in h_m \circ \mathrm{Sym}_\ell = [h_m]_\ell$. □

G.2 Orbit disjointness

Proposition 20 handles collisions within a fixed component. The following condition extends orbit disjointness to *cross-component* collisions, where a single branch matches different truth branches on different components.

Condition (OD) (Cross-component orbit disjointness). A pair (h, Y) satisfies (OD) if for distinct branches $j \neq m$ and distinct components $\ell \neq n$, $[h_j]_\ell \cap [h_m]_n = \varnothing$.

A violation means the difference map $h_m^{-1} \circ h_j$ lies in $\mathrm{Sym}_n \cdot \mathrm{Sym}_\ell^{-1}$, a set of dimension at most $2D$ inside $\mathrm{Aff}(\mathbb{R}^d)$, hence codimension at least $2d$.

See the algorithmic-information viewpoint discussion in Appendix F for a unified dimension-counting analysis of (OD) and (TOD).

G.3 Distinct weights and joint symmetry triviality

Condition (DW) (Distinct weights). A GMM has *distinct weights* if $w_k \neq w_\ell$ for all $k \neq \ell$.

Under (DW), each weight w_k acts as a unique scalar "tag" that sticks to component k through pushforward. In the global bijection Ψ from Lemma 2, coefficient matching forces $w_k = w'_\ell$; (DW) then forces $k = \ell$, making the bijection *component-preserving*.

Condition (JST) (Joint symmetry triviality). A GMM satisfies (JST) if $\bigcap_{k=1}^{K} \mathrm{Sym}_k = \{\mathrm{Id}\}$.

(JST) forbids nontrivial affine maps that fix every component individually. But crucially, it does *not* forbid maps that *permute* components—a key distinction from (MST).

Proposition 21. *The following implications hold: (1)* $(DW) + (JST) \implies (MST)$*; (2)* $(MST) \implies (JST)$*; (3) under* (DW)*,* $(JST) \iff (MST)$*.*

Proof. **(1)** Let $T_\sharp p = p$. Since T preserves the mixture law, it induces a permutation σ with $T_\sharp \mathcal{N}_k = \mathcal{N}_{\sigma(k)}$ and $w_k = w_{\sigma(k)}$ for all k. Under (DW), distinct weights force $\sigma = \mathrm{id}$, so $T \in \bigcap_k \mathrm{Sym}_k$, and (JST) gives $T = \mathrm{Id}$. **(2)** If $T \in \bigcap_k \mathrm{Sym}_k$, then $T_\sharp p = p$; (MST) gives $T = \mathrm{Id}$. **(3)** Immediate from (1) and (2). □

Remark G.1 (The (JST)–(MST) gap). Without (DW), (JST) and (MST) can diverge. Consider $\mathcal{N}_1 = \mathcal{N}(\mu_1, \sigma^2)$, $\mathcal{N}_2 = \mathcal{N}(\mu_2, \sigma^2)$, with $\mu_1 \neq \mu_2$ and equal weights $w_1 = w_2 = \frac{1}{2}$. Then $\mathrm{Sym}_1 \cap \mathrm{Sym}_2 = \{\mathrm{Id}\}$: (JST) holds. But the reflection $T(x) = -x + (\mu_1 + \mu_2)$ swaps the two components; with equal weights, $T_\sharp p = p$, so (MST) fails. If a decoder branch is post-composed with T, the output distribution is unchanged but the map has been modified. (JST) cannot detect this switcheroo; (MST) can. This is why our main-text MID result uses (MST) rather than (JST). A two-dimensional extension of this example appears in Example 5 (Appendix K).

H Cycle determinant lemma

Lemma 22 (Cycle determinant)**.** *Let $p = \sum_{k=1}^{K} w_k \mathcal{N}(\mu_k, \Sigma_k)$ be an irredundant GMM, and let $T(x) = Ax + b$ be an affine symmetry of p, inducing a permutation $\sigma \in S_K$ of the components. Then on each cycle of σ, the covariance determinants are constant.*

In particular, if the numbers $\det \Sigma_1, \ldots, \det \Sigma_K$ are pairwise distinct, then $\sigma = \mathrm{id}$.

Proof. If T sends component k to component $\sigma(k)$, then $A\Sigma_k A^\top = \Sigma_{\sigma(k)}$. Hence

$$\det \Sigma_{\sigma(k)} = (\det A)^2 \det \Sigma_k.$$

Now follow a cycle $k \to \sigma(k) \to \cdots \to \sigma^{\ell-1}(k) \to k$. Multiplying the determinant relations gives

$$\det \Sigma_k = (\det A)^{2\ell} \det \Sigma_k.$$

Since $\det \Sigma_k > 0$, this implies $(\det A)^{2\ell} = 1$, hence $(\det A)^2 = 1$. Therefore

$$\det \Sigma_{\sigma(k)} = \det \Sigma_k$$

for every k, and determinants are constant along each cycle. □

I Sufficient conditions for symmetry triviality

We present four sets of sufficient conditions for (MST), arising from the combinations $\{(DW),\ \text{distinct } \det(\Sigma)\} \times \{\text{affinely independent means},\ \text{2-comp generic position}\}$.

I.1 Reducing (MST) to (JST)

Two sufficient conditions reduce (MST) to (JST):

1. **(DW) (distinct weights).** Under (DW), any affine symmetry must fix every component individually (Proposition 21), so (MST) reduces to (JST).
2. **Distinct covariance determinants.** By Lemma 22, if $\det \Sigma_1, \ldots, \det \Sigma_K$ are pairwise distinct, any affine symmetry induces $\sigma = \mathrm{id}$, again reducing (MST) to (JST).

I.2 Sufficient conditions for (JST)

Two sufficient conditions for (JST):

1. **Affinely independent means.**
   Lemma 23 (Affinely independent means ⇒ (JST))**.** *If the means $\{\mu_k\}_{k=1}^K$ are in affinely general position (i.e., the direction space $U := \operatorname{span}\{\mu_k - \mu_1 : k = 2, \dots, K\}$ has dimension d), then the GMM satisfies (JST).*

   Proof. Let $T(x) = Ax + b \in \bigcap_k \operatorname{Sym}_k$. Then T fixes each component, so $A\Sigma_k A^\top = \Sigma_k$ for every k and $b = (I - A)\mu_k$ for every k. The second condition gives $(I - A)(\mu_k - \mu_j) = 0$ for all j, k, implying $A|_U = I|_U$. If $\dim U = d$, then $A = I$ and $b = 0$. □

   Note that this condition requires $K \geq d + 1$ components.
2. **2-component generic position.**
   Theorem 24 (Generic affine MST for GMMs)**.** *Let $M = \sum_{i=1}^K w_i \mathcal{N}(\mu_i, \Sigma_i)$, $K \geq 2$, be an ordered irredundant GMM on $\mathbb{R}^d$. Suppose:*
   (a) the matrix $H := \Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}$ has simple spectrum;
   (b) in an eigenbasis of H, the vector $v := \Sigma_1^{-1/2}(\mu_2 - \mu_1)$ has all coordinates nonzero.
   Then every element of $\bigcap_{k=1}^K \operatorname{Sym}_k$ is the identity. Therefore (JST) holds on an open dense full-measure subset of GMM parameter space.

   Proof. Let $x \mapsto Ax + b$ fix every component individually. In particular it fixes components 1 and 2. Set $O := \Sigma_1^{-1/2} A \Sigma_1^{1/2}$. Then $O \in O(d)$, and $b = (I - A)\mu_1$. Fixing the second component gives $OHO^\top = H$ and $Ov = v$. Since $O^{-1} = O^\top$, the first identity implies $OH = HO$. Because H has simple spectrum, every orthogonal matrix commuting with H is diagonal with entries ± 1 in an eigenbasis of H. In that basis, v has no zero coordinate; therefore $Ov = v$ forces every sign to be $+1$. Hence $O = I$, so $A = I$ and $b = 0$.
   The two conditions define an open dense full-measure subset: simple spectrum excludes the discriminant-zero hypersurface; the nonzero-coordinate condition excludes finitely many hyperplanes in v-space on the simple-spectrum locus. □

I.3 Summary of four combinations

Combining the two reduction routes with the two (JST) conditions gives four sets of sufficient conditions for (MST):

Reduce to JST via	JST via	Practicality
(DW)	affinely indep. means	requires $K > d$; DW often violated
(DW)	2-comp generic position	DW often violated
distinct $\det(\Sigma)$	affinely indep. means	requires $K > d$
distinct $\det(\Sigma)$	2-comp generic position	**most applicable**

The first row is simple and practical: distinct weights are almost-sure under continuous weight priors, and affinely independent means hold once $K \geq d+1$ with generic means. The combination "distinct $\det(\Sigma)$ + 2-component generic position" is the most broadly applicable: (DW) is sometimes violated in statistical practice (equal-weight mixtures), while affinely independent means requires $K > d$. All four sufficient conditions are *generic*: each is violated only on a measure-zero algebraic subset of the parameter space.

Remark I.1 (Sufficient conditions for (MST)). Beyond the four explicit combinations above, sufficient conditions for (MST) include: (a) (DW) plus (JST) (Proposition 21); (b) general group-theoretic conditions extending to other exponential families; omitted here for scope. Condition (b) is the most general but least constructive.

J ICA ambiguity conditions

J.1 The (DD) condition

"Diagonal component covariances" $\neq$ "independent latent coordinates." These are different notions because mixtures couple coordinates through the shared component index. "Diagonal component covariances" means: for each mixture component $C = k$, $Z \mid (C = k) \sim \mathcal{N}(\mu_k, \Sigma_k)$ with Σ_k diagonal, so coordinates are independent *conditional on* C. "Independent latent coordinates" means the *unconditional* law factorizes: $p_Z(z) = \prod_{t=1}^{d} p_t(z_t)$. A generic GMM with diagonal Σ_k does *not* factorize like this. A clean diagnostic is the law of total covariance: for $s \neq t$, $\mathrm{Cov}(Z_s, Z_t) = \mathrm{Cov}\big(\mathbb{E}[Z_s \mid C], \mathbb{E}[Z_t \mid C]\big) + \mathbb{E}\big[\mathrm{Cov}(Z_s, Z_t \mid C)\big]$. If every Σ_k is diagonal, then $\mathrm{Cov}(Z_s, Z_t \mid C) = 0$, but the first term is generally nonzero. For example, in $\mathbb{R}^2$, $Z \sim \frac{1}{2}\mathcal{N}\big((0,0), I\big) + \frac{1}{2}\mathcal{N}\big((3,3), I\big)$ has diagonal Σ_k, yet $\mathrm{Cov}(Z_1, Z_2) = \frac{9}{4} > 0$, so (Z_1, Z_2) are not independent. Thus diagonal Σ_k is a *within-component* statement; coordinate-independence is a *mixture-level* statement. Corollary 11 proves an ICA-form affine ambiguity under diagonal component covariances; full classical ICA (independent coordinates) is a further specialization.

The following condition is borrowed from Kivva et al. [26].

Condition (DD) (Distinct diagonals). A GMM Y with parameters $\{(\mu_k, \Sigma_k)\}_k$ satisfies (DD) if (i) every covariance Σ_k is diagonal, and (ii) there exist indices k_1, k_2 such that the spectral ratios

$$\Big(\frac{(\Sigma_{k_1})_{tt}}{(\Sigma_{k_2})_{tt}}\Big)_{t=1}^{d}$$

are pairwise distinct across dimensions $t \in \{1, \ldots, d\}$.

J.2 Proof of Corollary 11

Proof. By Theorem 5 (resp. Theorem 8), there exists an affine bijection

$$\alpha(z) = Az + c, \qquad A \in \mathrm{GL}(d).$$

By the component matching in Theorem 5, there is a permutation σ of $[K]$ such that

$$A\Sigma'_\ell A^\top = \Sigma_{\sigma(\ell)}$$

for all $\ell \in [K]$.

Choose truth indices r, s given by condition (DD), so that the diagonal entries of

$$D := \Sigma_r \Sigma_s^{-1}$$

are pairwise distinct. Let

$$r' := \sigma^{-1}(r), \qquad s' := \sigma^{-1}(s), \qquad D' := \Sigma'_{r'}(\Sigma'_{s'})^{-1}.$$

Since all covariance matrices are diagonal, both D and D' are diagonal, and

$$D = AD'A^{-1}.$$

Let e_j be the j-th coordinate vector. Since D' is diagonal,

$$D'e_j = (D')_{jj} e_j,$$

hence

$$D(Ae_j) = AD'e_j = (D')_{jj} Ae_j.$$

So each column Ae_j is an eigenvector of D. Because D is diagonal with pairwise distinct diagonal entries, its eigenspaces are exactly the coordinate axes. Therefore each column of A has exactly one nonzero entry. Since A is invertible, these nonzero entries occur in distinct rows, so

$$A = P\Lambda$$

for some permutation matrix P and some diagonal invertible matrix Λ. Thus

$$\alpha(z) = P\Lambda z + c.$$

□

Remark J.1 (Comparison with linear ICA). In classical linear ICA ($f(z) = Mz$ with mixing matrix M), identifiability states that M is recoverable up to right-multiplication by a signed permutation matrix. Corollary 11 extends this to nonlinear PWA decoders and GMM priors: the "mixing map" is a PWA function; the latent distribution is a GMM; and the identifiability takes the same algebraic form $\alpha(z) = P\Lambda\, z+c$. The structural conditions ((AD), (USB), (MST)) replace the non-Gaussianity assumption of classical ICA.

J.3 Alternative route: iVAE assumption (iv) implies ICA ambiguity

iVAE's assumption (iv) with $k = 2$ (Gaussian latents with modulated mean and variance) forces the diagonal *precisions* of the truth components to span all diagonal directions: by the block-rank argument in the proof of Proposition 30, the $2d$ natural-parameter differences of (iv) already make the variance block span $\mathbb{R}^d$, with no independent generation of parameters. In particular $d + 1$ components have linearly independent diagonal precision differences, which we now show implies the ICA-form ambiguity. (A "component" in the ICA sense—a latent dimension—is distinct from a mixture component, i.e. a domain.)

Proposition 25 (Affinely independent diagonal covariances or precisions ⇒ ICA ambiguity)**.** *Let Z be an irredundant GMM with $K \geq 2$ components, all covariances diagonal. Suppose there exist $d+1$ components $k_0, k_1, \ldots, k_d$ such that* either *the covariance differences* $\mathrm{diag}(\Sigma_{k_j}) - \mathrm{diag}(\Sigma_{k_0})$ or *the precision differences* $\mathrm{diag}(\Sigma_{k_j}^{-1}) - \mathrm{diag}(\Sigma_{k_0}^{-1})$ *($j = 1, \ldots, d$) are linearly independent in $\mathbb{R}^d$. Then any affine bijection $\alpha(z) = Az + c$ satisfying $\alpha(Z') \sim Z$ (with Z' also having diagonal covariances) must have $A = P\Lambda$ for a permutation matrix P and diagonal invertible Λ.*

Proof. Because α is invertible affine and both mixtures are irredundant, standard mixture identifiability gives a permutation $\sigma \in S_K$ such that for each component index ℓ,

$$A\Sigma'_\ell A^\top = \Sigma_{\sigma(\ell)}.$$

Let $\ell_j := \sigma^{-1}(k_j)$ for $j = 0, 1, \ldots, d$, and define diagonal matrices

$$D'_j := \Sigma'_{\ell_j} - \Sigma'_{\ell_0}, \qquad j = 1, \ldots, d.$$

Then

$$AD'_jA^\top = A(\Sigma'_{\ell_j} - \Sigma'_{\ell_0})A^\top = \Sigma_{k_j} - \Sigma_{k_0},$$

which is diagonal for each j. The matrices $\Sigma_{k_j} - \Sigma_{k_0}$ are linearly independent in the d-dimensional space of diagonal matrices (by assumption), hence the D'_j are also linearly independent because $M \mapsto AMA^\top$ is an invertible linear map on matrices. Therefore $\{D'_j\}_{j=1}^d$ is a basis of the diagonal subspace.

By linearity, for every diagonal matrix D we can write $D = \sum_{j=1}^d a_j D'_j$, hence

$$ADA^\top = \sum_{j=1}^d a_j (AD'_jA^\top)$$

is diagonal. In particular, for each standard basis diagonal matrix $E_t := \mathrm{diag}(e_t)$,

$$AE_tA^\top \text{ is diagonal.}$$

But $AE_tA^\top$ equals $a_t a_t^\top$, where a_t is the t-th column of A. A rank-one matrix $a_t a_t^\top$ is diagonal iff a_t has at most one nonzero coordinate. Hence each column of A has at most one nonzero entry. Since A is invertible, each column has exactly one nonzero entry, and these nonzero entries occur in distinct rows. Thus $A = P\Lambda$ for a permutation matrix P and an invertible diagonal Λ.

This settles the covariance hypothesis. The precision hypothesis follows by the *same argument with* $\Sigma \mapsto \Sigma^{-1}$: inverting $A\Sigma'_\ell A^\top = \Sigma_{\sigma(\ell)}$ gives $\Sigma'^{-1}_\ell = A^\top \Sigma^{-1}_{\sigma(\ell)} A$, so with $\widetilde{D}_j := \Sigma^{-1}_{k_j} - \Sigma^{-1}_{k_0}$ one has $A^\top \widetilde{D}_j A = \Sigma'^{-1}_{\ell_j} - \Sigma'^{-1}_{\ell_0}$, again diagonal. Replacing $A(\cdot)A^\top$ by $A^\top(\cdot)A$ throughout, $A^\top E_t A = (A^\top e_t)(A^\top e_t)^\top$ is the rank-one outer product of the t-th *row* of A with itself; its being diagonal forces each row of A to have at most one nonzero entry, so again $A = P\Lambda$. □

The chain of implications is: iVAE assumption (iv) (Gaussian sources, modulated mean and variance, $K > 2d$ components) ⇒ $d + 1$ components with affinely independent diagonal *precisions* (by the block-rank argument of Proposition 30) ⇒ ICA ambiguity (Proposition 25).

K Worked examples

The five examples below, ordered by complexity, illustrate the distinct ways in which (AD), (SB)/(USB), and (MST) can fail, and the identifiability consequences that result. Throughout, (f, Z) is the *truth pair* and (g, Z') the *estimator pair*; we write $Z' \sim Z$ when both sides share the same latent law. Example 4 recapitulates Proposition 26 of Appendix K.1 in the unified framework; Example 5 extends the one-dimensional case of Remark G.1 to two dimensions.

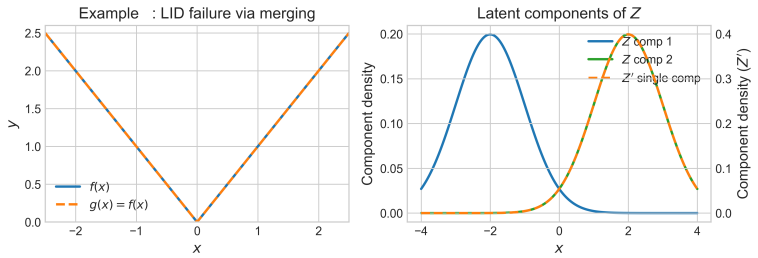


Figure 1: Example 1: $f = |x|$ collapses the symmetric two-component truth Z and the single Gaussian estimator Z' to the same observed law. LID fails; the two latents are not related by any affine map.

Example 1 (Fold-collapse; LID fails catastrophically). Setup. Take $f = g = |x|$ with branches $P_1 = [0, \infty)$, $x \mapsto x$ and $P_2 = (-\infty, 0)$, $x \mapsto -x$. Truth latent: $Z \sim \frac{1}{2}\mathcal{N}(-2, 1) + \frac{1}{2}\mathcal{N}(2, 1)$. Estimator latent: $Z' \sim \mathcal{N}(2, 1)$ (single component).

(SB)/(USB) for f. Both branches map into $(0, \infty)$, so $\mathcal{I}^f(y) = \varnothing$ for $y < 0$ and $\mathcal{I}^f(y) = \{1, 2\}$ for $y > 0$. At the only boundary $y = 0$, *both* branches enter simultaneously; no single branch toggles alone. Hence f violates (SB) and (USB).

(AD) for (f, Z). With slopes $A_1 = 1$ and $A_2 = -1$, the four branch-component pushforward parameters are

$$\theta^{(f)}_{1,1} = (-2, 1), \quad \theta^{(f)}_{1,2} = (2, 1), \quad \theta^{(f)}_{2,1} = (2, 1), \quad \theta^{(f)}_{2,2} = (-2, 1).$$

Pairs $(1, 1)$ and $(2, 2)$ collide at $(-2, 1)$; pairs $(1, 2)$ and $(2, 1)$ collide at $(2, 1)$. Hence (f, Z) violates (AD). For (f, Z') (single Gaussian), the two parameters $(-2, 1)$ and $(2, 1)$ are distinct, so (f, Z') satisfies (AD).

(MST). The reflection $T(x) = -x$ swaps the two components of Z with equal weights, so $G_{\text{mix}}(Z)$ is nontrivial: Z fails (MST). Any single Gaussian $\mathcal{N}(\mu, \sigma^2)$ is preserved by $x \mapsto 2\mu - x$, so Z' likewise fails (MST).

Equal pushforwards. For any absolutely continuous X with density p_X, the density of $|X|$ satisfies $p_{|X|}(y) = p_X(y) + p_X(-y)$ for $y > 0$. Using $\varphi(y; \mu, \sigma^2) = \varphi(-y; -\mu, \sigma^2)$ one computes $p_{|Z|}(y) = p_{|Z'|}(y)$ for all $y > 0$, so $f(Z) \sim f(Z')$.

LID fails. Any affine image of the Gaussian Z' is again Gaussian, hence cannot equal the genuine two-component mixture Z: no affine bijection α satisfies $\alpha(Z') \sim Z$.

Condition summary for Example 1:

	Truth (f, Z)	Estimator (f, Z')
(AD)	×	✓
(SB)	×	×
(USB)	×	×
(MST)	×	×

Example 2 (Kivva D.2; LID holds, MID fails). Setup. This is Example D.2 of Kivva et al. [26]. Let $Z \sim \frac{1}{2}\mathcal{N}(-2, 1) + \frac{1}{2}\mathcal{N}(2, 1)$ and $Z' \sim Z$ (shared law). The PWA maps $f, g \colon \mathbb{R} \to \mathbb{R}$ are

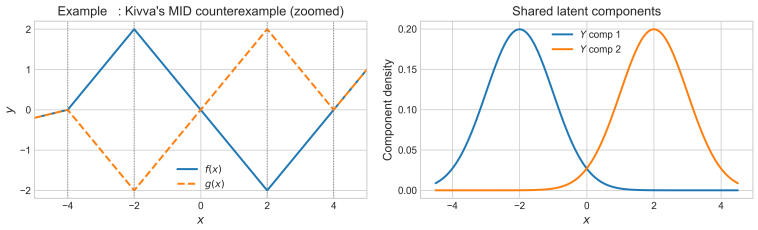


Figure 2: Example 2 (Kivva D.2): two different PWA maps f (top) and g (bottom) push forward the same shared latent Z to the same observed distribution. LID is trivial ($\alpha = \mathrm{Id}$), but MID fails because all truth-side conditions are violated.

defined by

$$f(x) = \begin{cases} x - 4, & x \geq 2, \\ -x, & -2 \leq x < 2, \\ x + 4, & -4 \leq x < -2, \\ (x + 4)/5, & x < -4, \end{cases} \qquad g(x) = \begin{cases} x - 4, & x \geq 4, \\ -x + 4, & 2 \leq x < 4, \\ x, & -2 \leq x < 2, \\ -x - 4, & -4 \leq x < -2, \\ (x + 4)/5, & x < -4. \end{cases}$$

(AD). Computing branch-component parameters $(A_i\mu_k + b_i,\ A_i^2)$ for the two components ($\mu_1 = -2$, $\mu_2 = 2$, unit variances): the branch $x \mapsto x + 4$ on $[-4, -2)$ gives $(2, 1)$ and $(6, 1)$, while the branch $x \mapsto -x$ on $[-2, 2)$ gives $(2, 1)$ and $(-2, 1)$. The parameter $(2, 1)$ is achieved by two branch-component pairs in f: (f, Z) violates (AD). For (g, Z'), the parameter $(-2, 1)$ is achieved by three separate branch-component pairs: (AD) fails even more severely.

(SB)/(USB). The chambers of both maps partition the observational axis at $y \in \{-2, 0, 2\}$. For f: at $y = 0$ exactly one branch enters (the branch mapping $[-4, -2)$, whose image is $(0, 2)$); this is a simple boundary. At $y = -2$ and $y = 2$, sets of two or three branches toggle together. Hence f satisfies (SB) but violates (USB). For g: no boundary has a unique toggling branch; g violates both (SB) and (USB).

(MST). The same reflection $T(x) = -x$ as in Example 1 is a nontrivial mixture symmetry of Z: (MST) fails.

Identifiability. LID holds trivially with $\alpha = \mathrm{Id}$ (shared latent law). MID fails: no affine α satisfies $f = g \circ \alpha^{-1}$.

Relation to our framework. Kivva concluded that *global injectivity* of f is necessary for MID; our conditions show instead that (AD)+(USB)+(MST) (latter two on the truth side) suffice—conditions this example violates entirely. Concretely, the difference map $S := f^{-1} \circ g$ (defined branch-wise using local invertibility) satisfies $S(Z) \sim Z$ but is a genuinely non-affine PWA map that violates (AD): the same latent-side non-affine symmetry pattern as in Example 1, and foreshadowing Example 3.

Condition summary for Example 2:

	Truth (f, Z)	Estimator (g, Z')
(AD)	×	×
(SB)	✓	×
(USB)	×	×
(MST)	×	×

Example 3 ("Example 101"; injective maps, MID fails). Setup. Let

$$f(x) = \begin{cases} -x, & |x| \leq 1, \\ x, & |x| > 1, \end{cases} \qquad g(x) = x, \qquad Z \sim \tfrac{1}{2}\mathcal{N}(-2, 1) + \tfrac{1}{2}\mathcal{N}(2, 1),$$

with $Z' \sim Z$ (shared law). Under the reduced-PWA definition (branch domains are finite unions of polyhedra), f has a *middle branch* $M\colon x \mapsto -x$ on $[-1, 1]$ and an *outer branch* $O\colon x \mapsto x$ on $(-\infty, -1) \cup (1, \infty)$. Both f and $g = \mathrm{Id}$ are injective.

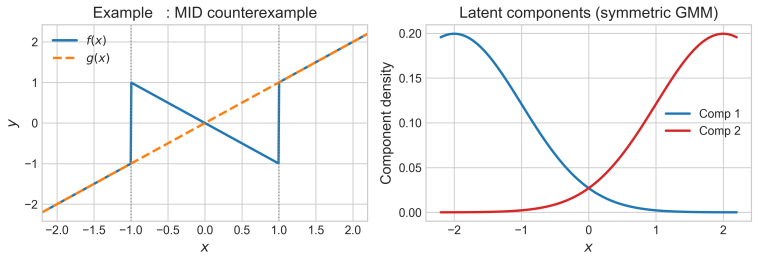


Figure 3: Example 3 ("Example 101"): both f and $g = \mathrm{Id}$ are injective, and f satisfies (USB), yet MID fails. The symmetric latent Z and the (AD) failure of (f, Z) conspire to hide the difference between the two maps.

Equal pushforwards. Since Z is symmetric under $x \mapsto -x$, the measure of Z on $[-1, 1]$ is invariant under the reflection performed by branch M. Combined with the identity action of branch O on the tails, $f(Z) \sim Z = g(Z')$.

(USB) for f. Branch images: $M \to [-1, 1]$, $O \to (-\infty, -1) \cup (1, \infty)$; three chambers $(-\infty, -1)$, $(-1, 1)$, $(1, \infty)$. At $y = -1$, M enters and O leaves (one branch each); at $y = 1$, M leaves and O enters. Both branches enjoy a simple boundary, so f satisfies (USB).

(AD) for (f, Z). Slopes: $A_M = -1$, $A_O = 1$. Branch-component parameters $(A_i\mu_k + b_i,\ A_i^2)$:

$$\theta^{(f)}_{M,1} = (2, 1), \quad \theta^{(f)}_{M,2} = (-2, 1), \quad \theta^{(f)}_{O,1} = (-2, 1), \quad \theta^{(f)}_{O,2} = (2, 1).$$

Pairs $(M, 1)$ and $(O, 2)$ both yield $(2, 1)$; pairs $(M, 2)$ and $(O, 1)$ both yield $(-2, 1)$. Hence (f, Z) violates (AD). For $(g, Z') = (\mathrm{Id}, Z)$, the parameters $(-2, 1)$ and $(2, 1)$ are distinct: (AD) holds.

(MST) for Z. Same as Examples 1–2: $T(x) = -x$ is a nontrivial mixture symmetry, so Z fails (MST).

MID fails. $f \neq g$ yet $f(Z) \sim g(Z')$: no affine α satisfies $f = g \circ \alpha^{-1}$. This is a genuinely injective MID counterexample—injectivity of f does not rescue identifiability when (AD) and (MST) both fail.

Condition summary for Example 3:

	Truth (f, Z)	Estimator (g, Z')
(AD)	×	✓
(SB)	✓	✓
(USB)	✓	✓
(MST)	×	×

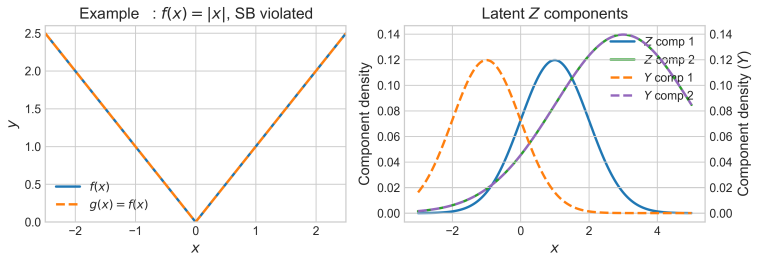


Figure 4: Example 4: the fold $f = |x|$ maps the truth Z and the one-component-reflected Y to the same observed law. (AD) and (MST) hold for both latents, but f fails (SB) and LID fails—showing (SB) cannot be omitted even when the other conditions are in force.

Example 4 (Single-flip fold; (AD) and (MST) hold, (SB) fails, LID fails). **Setup.** Let $f(x) = |x|$ (same (SB)/(USB) analysis as Example 1), and define

$$Z \sim \tfrac{1}{2}\mathcal{N}(1, 1) + \tfrac{1}{2}\mathcal{N}(3, 4), \qquad Y \sim \tfrac{1}{2}\mathcal{N}(-1, 1) + \tfrac{1}{2}\mathcal{N}(3, 4).$$

This is the setting of Proposition 26 (Appendix K.1); we recall the key points here.

Results (see Appendix K.1 for proofs).

1. $f(Z) \sim f(Y)$: the fold identifies $\mathcal{N}(1,1)$ with $\mathcal{N}(-1,1)$ in the pushforward, while the shared $\mathcal{N}(3,4)$ component is unchanged.
2. Both Z and Y satisfy (MST): the distinct variances (1 vs. 4) forbid component-swapping, and no single reflection fixes both components individually.
3. Both pairs (f,Z) and (f,Y) satisfy (AD): the four pushforward parameters $\mathcal{N}(\pm 1,1), \mathcal{N}(\pm 3,4)$ are pairwise distinct.
4. No affine bijection α satisfies $\alpha(Y) \sim Z$: distinct variances (1 vs. 4) forbid component-swapping, and neither $\alpha(x) = x + 2$ nor $\alpha(x) = -x + 2$ is consistent across both components simultaneously.

Hence LID fails despite (AD) and (MST) holding on both sides. The obstruction is entirely on the map side: $f = |x|$ has no simple boundary (point (1) of Appendix K.1's proof). This example belongs to a general *reflection-fold family* described in the remark following Proposition 26.

Condition summary for Example 4:

	Truth (f,Z)	Estimator (f,Y)
(AD)	✓	✓
(SB)	×	×
(USB)	×	×
(MST)	✓	✓

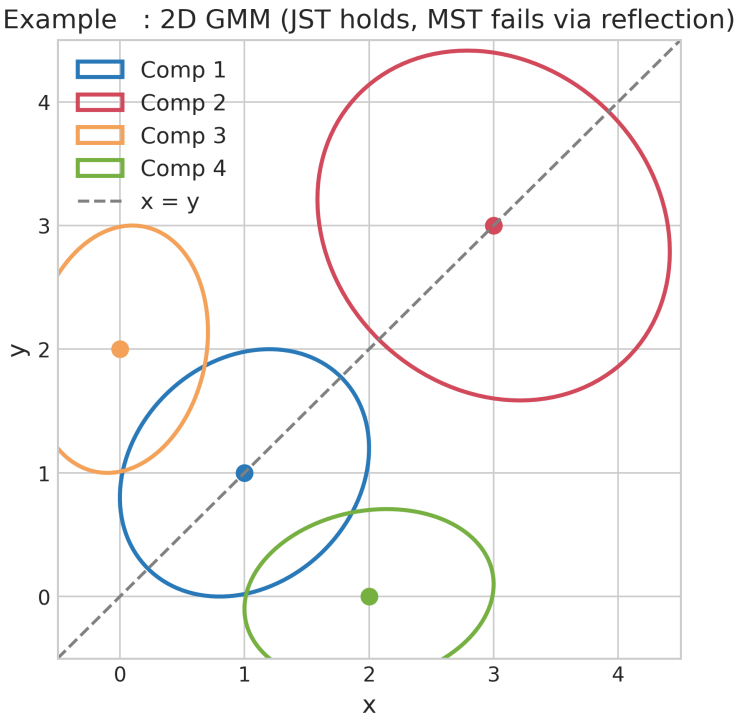


Figure 5: Example 5: a 2D four-component GMM satisfying (JST) but failing (MST). Components 1 and 2 lie on the diagonal $x_1 = x_2$, fixed individually by the swap $S(x_1,x_2) = (x_2,x_1)$. Components 3 and 4 are reflections of each other across this diagonal, swapped by S with equal weights $w_3 = w_4$. Equal weights cause the (JST)–(MST) gap; under (DW) the two conditions coincide (Proposition 21).

Example 5 ((JST) holds, (MST) fails; 1D and 2D). This example illustrates why the main-text MID results use (MST) rather than the strictly weaker (JST) (cf. Remark G.1).

One-dimensional case (Remark G.1). Let $p = \frac{1}{2}\mathcal{N}(\mu_1, \sigma^2) + \frac{1}{2}\mathcal{N}(\mu_2, \sigma^2)$ with $\mu_1 \neq \mu_2$. No single affine map can simultaneously reflect about both μ_1 and μ_2, so $\mathrm{Sym}_1 \cap \mathrm{Sym}_2 = \{\mathrm{Id}\}$ and (JST) holds. However, the reflection $T(x) = -x + (\mu_1 + \mu_2)$ swaps the two components with equal weights: $T_\sharp p = p$, $G_{\mathrm{mix}}(p) \neq \{\mathrm{Id}\}$, and (MST) fails. A decoder branch post-composed with T is undetectable from the observed law; (JST) cannot detect this switch, (MST) can.

Two-dimensional extension. Work in $\mathbb{R}^2$. Let $S(x_1, x_2) = (x_2, x_1)$ (the reflection across the diagonal $x_1 = x_2$, with matrix $P = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$). Define four components:

$$\mu_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 2 & -0.3 \\ -0.3 & 2 \end{pmatrix},$$

$$\mu_3 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 1 \end{pmatrix}, \qquad \mu_4 = P\mu_3 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \quad \Sigma_4 = P\Sigma_3 P^\top = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 0.5 \end{pmatrix},$$

with weights $w_1 = 0.2$, $w_2 = 0.3$, $w_3 = w_4 = 0.25$.

(MST) fails. For $k = 1, 2$: both means satisfy $\mu_{k,1} = \mu_{k,2}$ and both covariances satisfy $(\Sigma_k)_{11} = (\Sigma_k)_{22}$, so $S_\sharp\mathcal{N}(\mu_k, \Sigma_k) = \mathcal{N}(\mu_k, \Sigma_k)$. For $k = 3, 4$: by construction $S_\sharp\mathcal{N}(\mu_3, \Sigma_3) = \mathcal{N}(\mu_4, \Sigma_4)$ and vice versa. Since $w_3 = w_4$, we have $S_\sharp p = p$, so $G_{\mathrm{mix}}(p)$ contains the nontrivial element S and (MST) fails.

(JST) holds. Suppose $T(x) = Ax + b$ fixes all four components individually. From components 1 and 2 (means on the diagonal): $A(3,3)^\top + b = (3,3)^\top$ and $A(1,1)^\top + b = (1,1)^\top$, giving $A(1,1)^\top = (1,1)^\top$. From components 3 and 4: $A(2,0)^\top + b = (2,0)^\top$ and $A(0,2)^\top + b = (0,2)^\top$, giving $A(1,-1)^\top = (1,-1)^\top$. Since $(1,1)^\top$ and $(1,-1)^\top$ span $\mathbb{R}^2$, we obtain $A = I$ and $b = 0$: $\mathrm{Sym}_1 \cap \cdots \cap \mathrm{Sym}_4 = \{\mathrm{Id}\}$, so (JST) holds.

Condition	GMM p
(JST)	✓
(MST)	×
(DW)	× ($w_3 = w_4$)

K.1 A fold example showing necessity of boundary richness

The next example isolates a map-side obstruction. It shows that latent (MST) and (AD) do not suffice for law identifiability when the decoder has a fold and no (SB)-type boundary.

Proposition 26. *Let*

$$f(x) = |x| = \begin{cases} -x, & x < 0, \\ x, & x \geq 0, \end{cases}$$

and define the two latent laws

$$Z \sim \tfrac{1}{2}\mathcal{N}(1,1) + \tfrac{1}{2}\mathcal{N}(3,4), \qquad Y \sim \tfrac{1}{2}\mathcal{N}(-1,1) + \tfrac{1}{2}\mathcal{N}(3,4).$$

Then:

1. $f(Z) \sim f(Y)$.
2. *Both latent mixtures satisfy (MST).*
3. *Both pairs (f, Z) and (f, Y) satisfy (AD).*
4. *There is no affine bijection $\alpha \in \mathrm{Aff}(\mathbb{R})$ such that $\alpha(Y) \sim Z$.*

Consequently, law identifiability fails although latent (MST) and (AD) both hold.

Proof. **Step 1: equal pushforwards.** Because $f(x) = |x|$, $f_\sharp\mathcal{N}(1,1) = f_\sharp\mathcal{N}(-1,1)$. The second component $\mathcal{N}(3,4)$ is the same in both mixtures, so $f(Z) \sim f(Y)$.

Step 2: (SB) fails for f. The two branch images are both $[0, \infty)$. Hence the active set is $\mathcal{I}^f(x) = \varnothing$ for $x < 0$ and $\mathcal{I}^f(x) = \{-, +\}$ for $x > 0$. Across the boundary at 0, both branches enter simultaneously. So there is no boundary at which exactly one branch enters or leaves, and f fails (SB).

Step 3: latent (MST). We check Z; the argument for Y is identical. Let $T(x) = ax + b$ satisfy $T(Z) \sim Z$. Because the two component variances are 1 and 4, T cannot swap the two components: swapping would require simultaneously $a^2 \cdot 1 = 4$ and $a^2 \cdot 4 = 1$, which is impossible. So T must fix both components individually. A one-dimensional Gaussian $\mathcal{N}(\mu, \sigma^2)$ is fixed only by $x \mapsto x$ or $x \mapsto 2\mu - x$. To fix both $\mathcal{N}(1, 1)$ and $\mathcal{N}(3, 4)$ individually, T would have to be a reflection about 1 and about 3 at the same time, which is impossible. Hence $T = \mathrm{Id}$, and Z satisfies (MST).

Step 4: (AD). For (f, Z), the four branch-component pushforward Gaussians are $\mathcal{N}(1, 1)$, $\mathcal{N}(-1, 1)$, $\mathcal{N}(3, 4)$, $\mathcal{N}(-3, 4)$, all distinct. So (f, Z) satisfies (AD). For (f, Y), the four are $\mathcal{N}(-1, 1), \mathcal{N}(1, 1), \mathcal{N}(3, 4), \mathcal{N}(-3, 4)$, also all distinct.

Step 5: no affine latent reparametrization. Suppose $\alpha(x) = ax + b$ satisfies $\alpha(Y) \sim Z$. Because the component variances are 1 and 4, α cannot swap the two components. So it must send $\mathcal{N}(-1, 1) \mapsto \mathcal{N}(1, 1)$ and $\mathcal{N}(3, 4) \mapsto \mathcal{N}(3, 4)$. Variance preservation forces $a^2 = 1$.

If $a = 1$: the first condition gives $b = 2$, while the second gives $b = 0$. Contradiction.

If $a = -1$: the first condition gives $b = 0$, while the second gives $b = 6$. Contradiction.

So no such affine α exists. □

Remark K.1. This example is the one-dimensional member of a natural reflection-fold family. Let R be reflection across a hyperplane H, and define a two-branch map by $f(x) = A_+x + b$ on H^+ and $f(x) = A_+Rx + b$ on H^-. Then $f \circ R = f$. Replacing any latent Gaussian component G by its reflected image $R_\sharp G$ leaves its pushforward through f unchanged. For generic latent Gaussian parameters, this fold symmetry need not force latent (MST) or (AD) to fail; the obstruction is genuinely on the map side.

L Simple/universal boundaries vs. weak injectivity

This appendix clarifies how our boundary conditions (SB)/(USB) relate to global injectivity and to Kivva's *weak injectivity* [26]. All notation follows the main text.

L.1 Weak injectivity and 1-active chambers

Definition L.1 (Weak injectivity)**.** A PWA map $h\colon \mathbb{R}^d \to \mathbb{R}^d$ is *weakly injective* if there exists a nonempty open set $U \subset \mathbb{R}^d$ on which every $y \in U$ has a unique preimage.

Proposition 27. *Let h be a reduced PWA map with invertible branch rules. Then h is weakly injective if and only if some chamber C satisfies $|\mathcal{I}^h(C)| = 1$.*

Proof. ($\Rightarrow$) Let U be the open set given by weak injectivity. For $y \in U$, unique preimage forces $|\mathcal{I}^h(y)| = 1$. Since active sets are constant on each chamber, the chamber containing any $y \in U$ is 1-active.

($\Leftarrow$) If $\mathcal{I}^h(C) = \{i\}$, every $y \in C$ lies in $h(P_i)$ and in no other branch image. Invertibility of h_i gives a unique preimage on $U := C$. □

L.2 Global injectivity implies both USB and weak injectivity

Global injectivity forces every nonempty chamber to be 1-active, simultaneously yielding weak injectivity (trivially) and (USB) (Proposition 28). Global injectivity also implies weak injectivity, but the converse of both implications fails, as the remainder of this appendix shows.

L.3 Weak injectivity does not imply (SB)

We first identify the structural obstruction and then give a concrete example.

Remark L.1 (Co-active neighbouring branches). In a continuous 1-D PWA map, two branches sharing a domain breakpoint are forced to share the corresponding chamber boundary (by continuity $f(x_0^-) = f(x_0^+)$). If these two branches *both* cover an open interval on the same side of the boundary in observational space, they enter or leave together. A continuous map in which *every* pair of domain-neighbouring branches is co-active across *every* breakpoint therefore violates (SB)—no

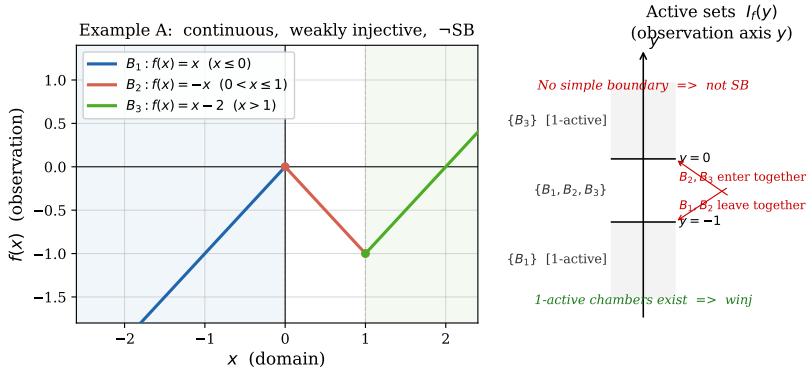


Figure 6: Example A: a continuous, weakly injective PWA map that violates (SB). *Left:* graph of f with three branches. The shaded regions indicate 1-active domains (which make the map weakly injective). *Right:* active-set diagram showing that at both boundaries, two branches toggle simultaneously.

boundary has a unique toggling branch. Any such map can be made weakly injective by attaching a tail branch with a single-branch 1-active chamber (e.g. a branch whose image extends to $\pm\infty$). The result is a map that is weakly injective yet violates (SB).

Example A (continuous, weakly injective, ¬SB). Define $f\colon \mathbb{R} \to \mathbb{R}$ by three branches on $(-\infty, 0]$, $(0, 1]$, $(1, \infty)$:

$$f(x) = \begin{cases} x, & x \le 0, \\ -x, & 0 < x \le 1, \\ x - 2, & x > 1. \end{cases}$$

This map is continuous: $f(0^-) = 0 = f(0^+)$ and $f(1^-) = -1 = f(1^+)$.

Branch images and chambers. $B_1 \to (-\infty, 0]$, $B_2 \to [-1, 0)$, $B_3 \to (-1, \infty)$. The observational axis splits into three chambers:

$$(-\infty, -1) : \ \mathcal{I}^f(\cdot) = \{B_1\}, \qquad (-1, 0) : \ \{B_1, B_2, B_3\}, \qquad (0, \infty) : \ \{B_3\}.$$

At $y = -1$, branches B_2 and B_3 enter together; at $y = 0$, branches B_1 and B_2 leave together. No boundary has a unique toggling branch, so f violates (SB) (and hence (USB)). On the other hand, chambers $(-\infty, -1)$ and $(0, \infty)$ are 1-active, so f is weakly injective (Figure 6).

Remark L.2 (Structural reading). Example A is a "weakly-injective tail attached to a co-toggling core": the middle portion (B_2, B_3) overlaps in image, forcing coupled toggles, while the outer tails produce the 1-active chambers. The same mechanism explains why g in Example 2 (Appendix K) violates (SB): every breakpoint boundary of g has co-active domain neighbours. We note that both f and g in Example 2 are weakly injective—the 1-active tail chambers $(-\infty, -2)$ and $(2, \infty)$ show this—yet g violates both (SB) and (USB).

Remark L.3 (Discontinuous maps: weak injectivity generically implies SB). The co-toggling mechanism of Remark L.1 requires continuity: domain neighbours must share image boundaries. Without continuity, a 1-active chamber C with $\mathcal{I}^f(C) = \{i\}$ has a boundary across which some branches enter. For a discontinuous map, these entering branches have image boundaries that are independent affine images of non-adjacent domain boundaries; having two or more of them coincide at the same $(d-1)$-dimensional boundary in observational space is a codimension-$(r-1)$ coincidence in the affine parameters (A_i, b_i). Thus, for generic discontinuous PWA maps, a 1-active chamber automatically yields a simple boundary, and weak injectivity implies (SB). This can be formalised by parametrising reduced PWA maps by their affine data: the set of maps where two non-neighbouring branch images share a boundary is a proper algebraic subvariety of the parameter space, hence measure-zero and nowhere-dense.

L.4 (SB) and (USB) do not imply weak injectivity

We give a discontinuous example and a continuous example, both satisfying (USB) while having no 1-active chamber.

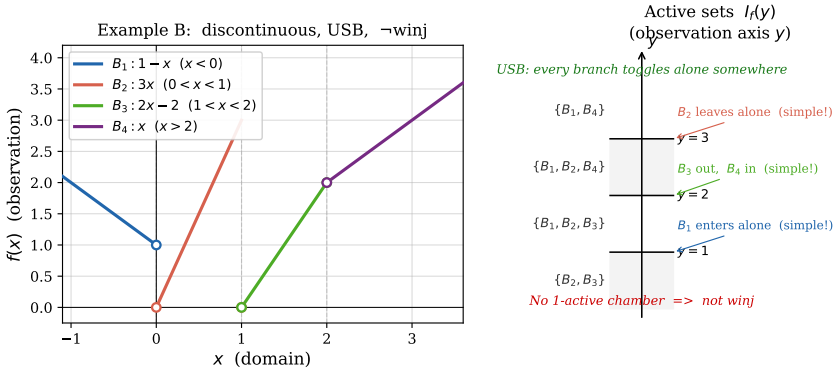


Figure 7: Example B: a discontinuous 4-branch PWA map satisfying (USB) with no 1-active chamber. *Left:* graph of f; open circles mark jump discontinuities. *Right:* every branch toggles alone at some boundary (USB), yet all chambers are 2-or-more-active (not weakly injective).

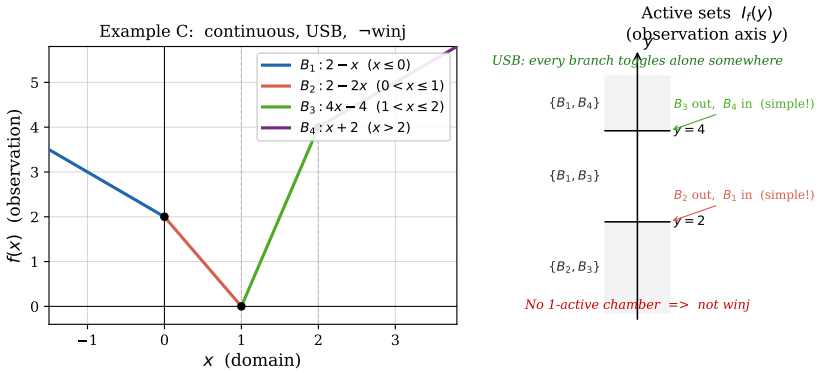


Figure 8: Example C: a continuous 4-branch PWA map satisfying (USB) with no 1-active chamber. *Left:* graph of f; filled dots mark continuity joints. *Right:* every branch toggles alone (USB), yet all three chambers are exactly 2-active.

Example B (discontinuous, USB, ¬winj). Let the branch domains be the open intervals $(-\infty, 0)$, $(0, 1)$, $(1, 2)$, $(2, \infty)$, with

$$f(x) = \begin{cases} 1 - x, & x < 0, \\ 3x, & 0 < x < 1, \\ 2x - 2, & 1 < x < 2, \\ x, & x > 2. \end{cases}$$

Branch images: $B_1 \to (1, \infty)$, $B_2 \to (0, 3)$, $B_3 \to (0, 2)$, $B_4 \to (2, \infty)$. The observational axis decomposes into four chambers:

$$(0, 1) : \{B_2, B_3\}, \quad (1, 2) : \{B_1, B_2, B_3\}, \quad (2, 3) : \{B_1, B_2, B_4\}, \quad (3, \infty) : \{B_1, B_4\}.$$

Every chamber has at least two active branches: f is *not* weakly injective. Yet (USB) holds: at $y = 1$, B_1 uniquely enters; at $y = 2$, B_3 uniquely leaves and B_4 uniquely enters; at $y = 3$, B_2 uniquely leaves. Every branch toggles alone at some boundary (Figure 7).

Example C (continuous, USB, ¬winj). Let the branch domains be $(-\infty, 0]$, $(0, 1]$, $(1, 2]$, $(2, \infty)$, with

$$f(x) = \begin{cases} 2 - x, & x \le 0, \\ 2 - 2x, & 0 < x \le 1, \\ 4x - 4, & 1 < x \le 2, \\ x + 2, & x > 2. \end{cases}$$

This map is continuous: $f(0) = 2$, $f(1) = 0$, $f(2) = 4$. Branch images: $B_1 \to [2, \infty)$, $B_2 \to [0, 2)$, $B_3 \to (0, 4]$, $B_4 \to (4, \infty)$. Chambers:

$$(0, 2) : \{B_2, B_3\}, \qquad (2, 4) : \{B_1, B_3\}, \qquad (4, \infty) : \{B_1, B_4\}.$$

Every chamber is exactly 2-active; f is not weakly injective. At $y = 2$, B_2 uniquely leaves and B_1 uniquely enters; at $y = 4$, B_3 uniquely leaves and B_4 uniquely enters. Each branch has a simple boundary: (USB) holds (Figure 8).

Remark L.4 (Genericity of the USB+$\neg$winj pattern). In both examples, the chamber structure is determined by the combinatorial type of the branch partition and the signs of the slopes. Small perturbations of slopes and intercepts (preserving continuity at breakpoints in Example C) yield the same sign patterns and hence the same chamber structure. Thus "USB $\wedge$ $\neg$winj" holds on an open set in the space of affine parameters: these examples are stable, not isolated.

L.5 (SB) does not imply global injectivity

(SB) is compatible with non-injective maps. For a continuous instance, the map f of Example 2 (Appendix K, originally Kivva et al. [26, Example D.2]) is continuous, weakly injective, and satisfies (SB), but is non-injective: in observational chamber $(-2, 0)$, three branches $\{A, C, D\}$ are simultaneously active. For discontinuous instances, Example B above satisfies (USB) (hence (SB)) while lacking even weak injectivity.

L.6 Summary of logical relations

Implication	**Reference**
global inj. $\Rightarrow$ (USB) $\Rightarrow$ (SB)	Prop. 28; definition
global inj. $\Rightarrow$ winj	trivial
winj $\not\Rightarrow$ (SB)	Example A (continuous); g in Ex. 2
(USB) $\not\Rightarrow$ winj	Examples B (discont.) and C (continuous)
(SB) $\not\Rightarrow$ winj	(follows from (USB) $\not\Rightarrow$ winj)
(SB) $\not\Rightarrow$ global inj.	f in Ex. 2; Example B

In summary, (USB) is strictly weaker than global injectivity, while (SB) and weak injectivity are logically incomparable. These conclusions hold identically in the continuous and discontinuous regimes. There are no further implications:

$$\text{global inj.} \implies (USB) \implies (SB), \qquad \text{global inj.} \implies \text{winj}, \qquad \text{no other arrows.} \tag{8}$$

L.7 Genericity: (SB)/(USB) versus weak injectivity

The examples above prove logical independence between (SB)/(USB) and weak injectivity. We separate two regimes. Under unconstrained (including discontinuous) parametrizations, (SB)/(USB) is generic (Remark L.5). Under a continuity constraint we have no proof and state a conjecture (Conjecture 1). The comparison with weak injectivity remains heuristic. Throughout, "non-generic" here means *failing on a set that—while sometimes of positive measure—shrinks as the branch count grows*, not the measure-zero notion used for (AD)/(MST) in the main text.

Remark L.5 (Discontinuous decoders). When discontinuities are allowed, (SB) is generic: a single non-co-toggling boundary suffices, and a generic PWA map has one. Its failure set (every boundary co-toggling) is of higher codimension than that of weak injectivity, so (SB) is strictly harder to violate. Likewise (USB) is generic: it fails only if every boundary of some branch is co-toggling. Nothing beyond the examples above is needed here.

Continuous decoders (open). When the decoder is constrained to be continuous we have no genericity proof for (SB). In one dimension (SB) can fail through an every-boundary folding construction; as this is a condition on slope signs it carves out a set of *positive* measure, so (SB) is not generic in the strict sense. But the construction needs at least two folds, so (SB) holds with high probability and its failure probability decays *exponentially in the branch count* M. The construction is one-dimensional; whether it extends to higher dimensions is open—co-toggling has more directions to exploit there—so we expect, but do not show, that a comparable bound survives.

Conjecture 1 (Continuous decoders)**.** *In every dimension d, the set of continuous PWA decoders that violate (SB), or (USB), shrinks as the branch count M grows.*

Weight-tied architectures are also open. They form a measure-zero slice of the space of continuous PWA decoders, so the conjecture would not cover them.

Weak injectivity is non-generic. Weak injectivity fails in both regimes, but the strength of the failure depends on the chamber type.

Bounded chambers (exponentially unlikely). Fix branch count M and pretend that on each chamber C the event "branch i is active" occurs independently with probability $p \in (0, 1)$. Then

$$\Pr(|\mathcal{I}^f(C)| = 1) = M\, p\, (1-p)^{M-1} \xrightarrow{M\to\infty} 0,$$

provided p stays bounded away from 0—equivalently, the average branch coverage grows with M, as when bounded branch images tile a fixed region ever more densely (a covering/packing argument). Under that condition 1-active chambers, hence weak injectivity, become exponentially rare.

Unbounded chambers (non-generic, but not exponentially rare). On unbounded tails the coverage need not grow, and the failure is milder. In a 1-D continuous PWA map the two unbounded branches have slopes a_L, a_R: if $\operatorname{sign}(a_L) = \operatorname{sign}(a_R)$ their images run to opposite infinities (1-active tails), otherwise to the same infinity (2-active tails). Under a symmetric random-sign model, tail weak injectivity already fails with probability $\approx \frac{1}{2}$—a *positive-measure* event, not a measure-zero coincidence. This unbounded case is the most favourable one for weak injectivity, and even here it is non-generic.

The upshot: weak injectivity demands a *global image-exclusivity* event—one branch dominating an open region unopposed—which is fragile, and exponentially so on bounded chambers. (SB)/(USB) demand only a *local combinatorial asymmetry* at a single boundary, which is robust and, once discontinuities are allowed, generic.

M Injectivity implies (USB) and chamber-wise (AD)

Proposition 28. *If $f : \mathbb{R}^d \to \mathbb{R}^d$ is a globally injective reduced PWA map, then f satisfies (USB).*

Proof. Under global injectivity, for any x in the interior of $\operatorname{Im}(f)$, there is exactly one z with $f(z) = x$. Hence (ignoring measure-zero boundary bookkeeping) every nonempty chamber has exactly one active branch: $|\mathcal{I}^f(C)| \in \{0, 1\}$.

Given any branch i, pick any chamber $C \subset f(P_i)^\circ$. If $f(P_i)^\circ = \mathbb{R}^d$ (the single-chamber case), clause (a) of (USB) applies. Otherwise, $\partial f(P_i) \neq \varnothing$, so there is an adjacent chamber C' across a boundary contained in $\partial f(P_i)$. Since chambers have constant active set and C' cannot have active set $\{i\}$ (it is across a boundary of branch i's image), the active set of C' is either $\varnothing$ or $\{i'\}$ with $i' \neq i$. In both cases, i is the unique leaving branch across (C, C'). So (USB) holds.

Caveat. A reduced branch domain P_i may be a union of disconnected polyhedra, hence $f(P_i)^\circ$ can have multiple connected components (multiple chambers) with the same singleton active set $\{i\}$. This does not break the argument. □

Proposition 29. *If f is globally injective, then (f, Z) satisfies chamber-wise (AD) for any irredundant GMM Z.*

Proof. On any chamber C with $\mathcal{I}^f(C) = \{i\}$, the active pairs are $\{i\} \times [K]$. Since f_i is an invertible affine map and the GMM is irredundant, the map $k \mapsto \theta^{(f)}_{i,k}$ is injective. Hence chamber-wise (AD) holds automatically. □

As shown in Appendix E, injectivity does *not* imply global (AD).

N Observed domain labels and iVAE assumption (iv) imply (MST)

We show that, when the PWA+GMM model is viewed as a multi-domain model with *observed* domain labels, iVAE's sufficient-variability condition (assumption (iv) of Khemakhem et al. [25]) implies our (MST)—or, once labels are observed, its labelled form (JST).

In iVAE, the latent distribution is a conditionally factorial exponential family with $n = d$ latent dimensions and k sufficient statistics per dimension. For Gaussian latents with modulated mean and variance, $k = 2$, so the natural parameter vector $\lambda(u) \in \mathbb{R}^{2d}$ has $nk = 2d$ entries. Assumption (iv) requires $nk + 1 = 2d + 1$ distinct domain values $u_0, \dots, u_{2d}$ such that the $2d \times 2d$ matrix $L = (\lambda(u_1) - \lambda(u_0), \dots, \lambda(u_{2d}) - \lambda(u_0))$ is invertible. In our GMM framework, each domain value corresponds to a mixture component, so assumption (iv) requires $K > 2d$ components. Each component is $\mathcal{N}(\mu(u), \Sigma(u))$ with $\Sigma(u)$ diagonal, and for the sufficient statistic $T(z_i) = (z_i, z_i^2)$ the natural parameter splits as $\lambda(u) = (\eta(u), \gamma(u)) \in \mathbb{R}^{2d}$ with $\eta(u) = \Sigma(u)^{-1}\mu(u)$ and $\gamma(u) = -\frac{1}{2}\operatorname{diag}(\Sigma(u)^{-1})$.

Observed domain labels turn the mixture into a *labelled GMM* in practice: any symmetry T with $T_\sharp p = p$ must preserve each label—because T maps the data from labelled component k to data that must still carry label k—so the permutation σ induced by T must be the identity. Consequently, admissible symmetries are precisely those that fix every component individually, i.e., $T \in \bigcap_k \mathrm{Sym}_k$. This reduces (MST) to (JST).

Proposition 30. *Suppose the GMM has $K > 2d$ components with diagonal covariances, the domain labels are observed, and iVAE assumption (iv) holds. Then the GMM satisfies (MST) (precisely, its label-preserving form (JST), which coincides with (MST) once labels are observed).* No independent generation of the mean and covariance parameters is required.

Proof. Since domain labels are observed, any affine self-symmetry fixes every component individually, reducing (MST) to (JST). It thus suffices to show that the $2d+1$ components singled out by (iv) have trivial common symmetry group, $\bigcap_{r=0}^{2d} \mathrm{Sym}_{u_r} = \{\mathrm{Id}\}$.

Let $T(x) = Ax + b$ fix each of these Gaussians, and write $\mu_r := \mu(u_r)$, $\Sigma_r := \Sigma(u_r)$ (diagonal). Fixing a nondegenerate Gaussian is equivalent to

$$A\Sigma_r A^\top = \Sigma_r, \qquad A\mu_r + b = \mu_r,$$

so $b = (I - A)\mu_r$ and, inverting the covariance equation, $A^\top \Sigma_r^{-1} A = \Sigma_r^{-1}$ for every r.

Step 1: (iv) spans the diagonal precisions. Split L into mean/variance blocks $L = \binom{X}{Y}$ with $Y \in \mathbb{R}^{d \times 2d}$ collecting the rows $\gamma(u_r) - \gamma(u_0)$. Every column of L lies in $\mathbb{R}^d \oplus \mathrm{col}(Y)$, so $2d = \mathrm{rank}(L) \le d + \mathrm{rank}(Y)$, forcing $\mathrm{rank}(Y) = d$. Since $\gamma(u) = -\frac{1}{2}\operatorname{diag}(\Sigma(u)^{-1})$, the vectors $\operatorname{diag}(\Sigma_r^{-1}) - \operatorname{diag}(\Sigma_0^{-1})$ span $\mathbb{R}^d$; equivalently, $\{\Sigma_r^{-1}\}_{r=0}^{2d}$ spans the space of diagonal $d \times d$ matrices.

Step 2: A fixes every diagonal matrix. The linear map $D \mapsto A^\top D A$ fixes each Σ_r^{-1}, hence, by Step 1, every diagonal D: $A^\top D A = D$. Taking $D = E_j$ (the diagonal matrix with a single 1 in entry (j, j)), $A^\top E_j A = (A^\top e_j)(A^\top e_j)^\top$ is the rank-one outer product of the j-th row of A with itself; equating it to E_j forces that row to be $\pm e_j^\top$. Thus $A = \operatorname{diag}(\varepsilon_1, \dots, \varepsilon_d)$ with $\varepsilon_j \in \{\pm 1\}$.

Step 3: (iv) rules out sign flips. From $b = (I - A)\mu_r$, coordinate j reads $b_j = (1 - \varepsilon_j)(\mu_r)_j$ for all r. If $\varepsilon_j = -1$ then $(\mu_r)_j \equiv c$ is constant in r, whence $\eta(u_r)_j = (\Sigma_r^{-1})_{jj}\,(\mu_r)_j = -2c\,\gamma(u_r)_j$; subtracting the $r = 0$ relation makes row j of L equal to $-2c$ times row $d + j$, contradicting $\mathrm{rank}(L) = 2d$. Hence every $\varepsilon_j = +1$, so $A = I$ and $b = (I - A)\mu_0 = 0$, i.e. $T = \mathrm{Id}$. Therefore $\bigcap_r \mathrm{Sym}_{u_r} = \{\mathrm{Id}\}$, giving (JST) and, with observed labels, (MST). □

Remark N.1. Under iVAE assumption (iv) with $K > 2d$ components, one can also extract $d + 1$ components with affinely independent diagonal *precisions* (Proposition 25), yielding the ICA-form ambiguity—with no independent generation of parameters. Indeed the block-rank argument in the proof of Proposition 30 shows that the $2d$ natural-parameter differences of (iv) already force the variance block, i.e. the diagonal precisions, to span $\mathbb{R}^d$.

O Finite-order affine symmetries of GMMs

Proposition 31. *Let $p = \sum_{k=1}^{K} w_k \mathcal{N}(\mu_k, \Sigma_k)$ be an irredundant nondegenerate Gaussian mixture on $\mathbb{R}^d$. If p admits a nontrivial affine self-symmetry, then it admits a nontrivial affine self-symmetry of finite order.*

Proof. Let $T(x) = Ax + b$ be a nontrivial affine map with $T_\sharp p = p$. Since affine maps send Gaussians to Gaussians and the Gaussian components are irredundant, uniqueness of finite Gaussian-mixture decomposition implies that T permutes the weighted components. Let $\sigma \in S_K$ be the induced permutation, and let m be its order.

Then T^m fixes every component individually. Write $T^m(x) = Cx + d$, $C = A^m$. If $T^m = \mathrm{Id}$, then T itself has finite order, and we are done.

Assume now that $T^m \neq \mathrm{Id}$. Because T^m fixes each component $\mathcal{N}(\mu_k, \Sigma_k)$, we have

$$C\Sigma_k C^\top = \Sigma_k, \qquad C\mu_k + d = \mu_k \qquad (k = 1, \dots, K).$$

Subtracting the mean equations for k and 1 gives $C(\mu_k - \mu_1) = \mu_k - \mu_1$ for all k.

Define

$$H := \big\{ L \in \mathrm{GL}(d) : L\Sigma_k L^\top = \Sigma_k,\ L(\mu_k - \mu_1) = \mu_k - \mu_1\ \forall k \big\}.$$

This is a closed subgroup of $O(\Sigma_1) := \{L \in \mathrm{GL}(d) : L\Sigma_1 L^\top = \Sigma_1\}$, hence H is a compact Lie group. Since $C \in H$ and $T^m \neq \mathrm{Id}$, we must have $C \neq I$, so H is nontrivial.

Every nontrivial compact Lie group contains a nontrivial finite-order element. Choose $L \in H$, $L \neq I$, of finite order. Then the affine map $S(x) := L(x - \mu_1) + \mu_1$ has finite order and fixes every Gaussian component:

$$S_\sharp \mathcal{N}(\mu_k, \Sigma_k) = \mathcal{N}(\mu_k, \Sigma_k) \qquad (k = 1, \dots, K).$$

Hence $S_\sharp p = p$, and S is a nontrivial finite-order affine self-symmetry of p. □

Remark O.1. This is the group-theoretic reason the (MST) assumption is structurally necessary for Gaussian-mixture latent laws in broad piecewise decoder classes: once a nontrivial affine symmetry T exists, a finite-order one S exists, and S can be localised into a two-piece strict PWA self-map $\{S, \mathrm{Id}\}$ (with appropriately chosen branch domains) that preserves the mixture law while genuinely modifying the decoder (detail and proof omitted for scope; see also Remark 4.2). The next section shows analogous torsion-based necessity phenomena for broad classes of exponential-family and location-scale mixtures.

P Broad torsion and symmetry-triviality mechanisms

The present paper uses only the Gaussian case. For orientation, Table 5 records several broader mechanisms that also arise for exponential-family and location-scale mixtures. Only the first row is proved here (Proposition 31); the remaining rows summarise complementary omitted results and directions of ongoing work.

Remark P.1. The broad message is that, beyond the Gaussian case, two phenomena recur: (1) *torsion-guaranteed settings*, where any nontrivial symmetry automatically produces a finite-order obstruction; (2) *symmetry-trivial settings*, where irredundant mixtures have no nontrivial symmetry at all. Both support the same conclusion: (MST) is not an artificial assumption, but the natural boundary between identifiable and non-identifiable regimes.

Q Alternative map identifiability theorems

The main-text MID theorem (Lemma 7 and Theorem 8) uses the boundary-richness condition (USB) to establish a branch bijection, and then (MST) to collapse the branch-wise ambiguities. We now present two alternative routes that replace (USB) with purely algebraic conditions.

Table 5: Broad torsion and symmetry-triviality mechanisms.

Setting	What happens	Mechanism	Examples
Irredundant Gaussian mixtures	Nontrivial symmetry $\Rightarrow$ finite-order element	compact stabilizer	proved here
Nondegenerate location-scale mixtures	likewise, finite order somewhere	compact linear stabilizer	Gaussian, Laplace, Student-t
Compact-group-preserving families	torsion already present	compact Lie group contains finite-order elements	Bingham-type
Countable-support families	involution already present	transposition inside nontrivial fiber	discrete EF
One-parameter dilation families	no nontrivial realised symmetry	$c^\ell = 1$ with $c > 0 \Rightarrow c = 1$	exponential, fixed-shape gamma
Trivial realised symmetry group	MST holds identically	no symmetry to begin with	Poisson

Q.1 MID via (DW), (OD), and (JST)

Theorem 32 (MID via DW/OD/JST)**.** *Let f, g be PWA maps and Z, Z' both follow the same GMM p with K components. Assume* (1)*, (AD) on both sides, and on the truth side: (OD) on (f, Z), (DW) and (JST) on p. Then $f = g$.*

Proof. **Step 1.** Since Z and Z' share law p with distinct weights ((DW)), the permutation aligning their components is unique; fix matched indexing.

Step 2. Coefficient matching requires $w_k = w'_\ell$; (DW) forces $k = \ell$. Hence for each (j, k) active on C, there is a unique $i_k(j)$ with $\theta^{(g)}_{j,k} = \theta^{(f)}_{i_k(j),k}$.

Step 3. We show $i_k(j)$ is independent of k. Suppose $k \neq \ell$ give $i' = i_k(j) \neq i'' = i_\ell(j)$. Then $g_j \in [f_{i'}]_k \cap [f_{i''}]_\ell$, which contradicts (OD). Hence $i_k(j) =: \pi(j)$ is constant in k. Injectivity of π follows from (AD); since $|I_f| = |I_g|$, π is a bijection.

Step 4. Fix j, set $i = \pi(j)$. Then $S_j := f_i^{-1} \circ g_j$ satisfies $S_j \in \bigcap_k \mathrm{Sym}_k$. By (JST), $S_j = \mathrm{Id}$, so $f_i = g_j$. Active-set agreement gives $f = g$. □

Remark Q.1 (Role of each condition).
- (DW) ensures component-preserving matching (Step 2), enabling the orbit argument. Without it, Ψ can swap components across branches, and (OD) becomes inapplicable.
- (OD) forces the branch-wise component match to be consistent across components (Step 3). Without it, different components could match to different truth branches.
- (JST) kills the residual stabiliser ambiguity (Step 4). Under (DW), this is equivalent to (MST) (Proposition 21).

Q.2 Transporter-orbit disjointness

For Gaussians $\mathcal{N}_k, \mathcal{N}_\ell$ of a GMM, define $\mathrm{Aff}(k{\to}\ell) := \{T \in \mathrm{Aff}(\mathbb{R}^d) : T_\sharp \mathcal{N}_k = \mathcal{N}_\ell\}$. This is the set of all affine maps that transport component k exactly onto component ℓ. It is a coset of the stabiliser: $\mathrm{Aff}(k{\to}\ell) = T_{k\to\ell} \circ \mathrm{Sym}_k$ for any fixed $T_{k\to\ell} \in \mathrm{Aff}(k{\to}\ell)$. Consequently, $\dim \mathrm{Aff}(k{\to}\ell) = D$. When $k = \ell$, the transporter set reduces to the stabiliser: $\mathrm{Aff}(k{\to}k) = \mathrm{Sym}_k$.

The *transporter-orbit* of branch i under the pair (k, ℓ) is

$$h_i \circ \mathrm{Aff}(k{\to}\ell) = \{h_i \circ T : T \in \mathrm{Aff}(k{\to}\ell)\}.$$

This is the set of all affine maps S such that branch i followed by S^{-1} transports $\mathcal{N}_k$ to $\mathcal{N}_\ell$.

Condition (TOD) (Transporter-orbit disjointness). A pair (h, Y) satisfies (TOD) if for all distinct branches $i \neq i'$ and all component quadruples (k, ℓ, k', ℓ'), $\big(h_i \circ \mathrm{Aff}(k{\to}\ell)\big) \cap \big(h_{i'} \circ \mathrm{Aff}(k'{\to}\ell')\big) = \varnothing$.

Relationship to other conditions. Setting $k = \ell$ and $k' = \ell'$ reduces $\mathrm{Aff}(k{\to}k) = \mathrm{Sym}_k$, so (TOD) implies (OD) (further restricted to $k \neq k'$) and SAMECOMPAD (further restricted to $k = k'$). (TOD) is thus strictly stronger than both.

Genericity. A (TOD) violation requires the difference map $h_{i'}^{-1} \circ h_i$ to lie in $\mathrm{Aff}(k' \to \ell')^{-1} \circ \mathrm{Aff}(k \to \ell)$, a set of dimension at most $2D$ (two independent orthogonal degrees of freedom, one from each covariance). The full affine group has dimension $d^2 + d$, so the bad set has codimension at least $2d$—the same bound as for (OD).

Q.3 MID via (TOD) and (MST)

Theorem 33 (MID via TOD/MST)**.** *Let f, g be PWA maps and Z, Z' both follow the same GMM p with K components. Assume* (1)*, (AD) on both sides, and on the truth side: (TOD) on (f, Z) and (MST) on p. Then $f = g$.*

Proof. **Step 1: Branch bijection via (TOD).** The global bijection $\Psi(j,k) = (i(j,k),\, \ell(j,k))$ from Lemma 2 satisfies $\theta^{(g)}_{j,k} = \theta^{(f)}_{i(j,k),\ell(j,k)}$ and $w_k = w_{\ell(j,k)}$. This means $g_j^{-1} \circ f_{i(j,k)}$ transports $\mathcal{N}_k$ to $\mathcal{N}_{\ell(j,k)}$, i.e.,

$$g_j \;\in\; f_{i(j,k)} \circ \mathrm{Aff}(k \to \ell(j,k)).$$

Fix any estimator branch j. Suppose there exist $k_1 \neq k_2$ with $i_1 := i(j,k_1) \neq i_2 := i(j,k_2)$. Then

$$g_j \;\in\; \big(f_{i_1} \circ \mathrm{Aff}(k_1 \to \ell_1)\big) \;\cap\; \big(f_{i_2} \circ \mathrm{Aff}(k_2 \to \ell_2)\big),$$

where $\ell_s := \ell(j,k_s)$. Since $i_1 \neq i_2$, this contradicts (TOD) on (f, Z). Hence $i(j,k) =: \pi(j)$ is independent of k.

Step 2: π is a bijection. If $\pi(j_1) = \pi(j_2)$ for $j_1 \neq j_2$, then Ψ maps both blocks $\{j_1\} \times [K]$ and $\{j_2\} \times [K]$ into $\{\pi(j_1)\} \times [K]$, requiring $2K \leq K$: impossible. Since $|I_f| = |I_g|$, π is a bijection.

Step 3: Kill branch differences via (MST). For each j, define $S_j := f_{\pi(j)}^{-1} \circ g_j$. The matching gives $S_{j\sharp}\mathcal{N}_k = \mathcal{N}_{\ell(j,k)}$ for every k, so S_j induces a bijection $\tau_j \colon k \mapsto \ell(j,k)$ on components. The weight equality $w_k = w_{\tau_j(k)}$ from the coefficient matching ensures this permutation preserves weights. Hence

$$\begin{aligned} S_{j\sharp}\, p &= \sum_k w_k\, \mathcal{N}_{\tau_j(k)} \\ &= \sum_m w_{\tau_j^{-1}(m)}\, \mathcal{N}_m = \sum_m w_m\, \mathcal{N}_m = p, \end{aligned}$$

so $S_j \in G_{\mathrm{mix}}(p)$. By (MST), $S_j = \mathrm{Id}$; hence $f_{\pi(j)} = g_j$. Active-set agreement gives $f = g$. □

Q.4 Comparison of approaches

Table 6: Three alternative routes to map identifiability (shared latent setting). All assume (AD) on both sides.

Route	Branch bij.	Kill
USB + MST	Boundary richness	MST
DW + OD + JST	DW (comp-pres.) + OD	JST (= MST)
TOD + MST	TOD	MST

In all three cases, once the branch bijection is established, the residual branch-wise ambiguity S_j lies in $G_{\mathrm{mix}}(p)$, and (MST) (or its equivalent under (DW)) forces $S_j = \mathrm{Id}$. This "two-step" structure—first a branch bijection, then a symmetry collapse—is a recurring pattern across all our identifiability results.

The three routes reflect different trade-offs:

- **USB + MST** (Lemma 7) uses geometric boundary structure to isolate each branch, and a single symmetry condition to kill the ambiguity. It requires no algebraic conditions on the map beyond (AD).
- **DW + OD + JST** (Theorem 32) uses weight distinctness as combinatorial tags and orbit geometry to separate branches. It requires no boundary conditions. Under DW, (JST) and (MST) coincide (Proposition 21), so the "killing" mechanism is equivalent.

- **TOD + MST** (Theorem 33) operates without boundary conditions or weight distinctness, using the generalised transporter-orbit structure to handle arbitrary component-branch entanglement. The cost is a stronger orbit condition ((TOD) vs. (OD)).

R Stability of map identifiability

The conditions of Theorem 8 are open: two atoms can drift together, an exposing boundary can shrink to nothing, a mixture can become symmetric in the limit. With unboundedly many regions no stability statement holds across all PWA decoders. So we fix a complexity budget κ and ask each condition to hold with a margin $c > 0$.

Notation. W_1 is the Wasserstein-1 distance, vol is Lebesgue measure, $\sigma(A)$ is the set of singular values of A, $\preceq$ is the positive-semidefinite order, and $B(x, r)$ is the open ball of centre x and radius r. Two windows of radius R are used: $B_R := B(0, R)$ in latent space, where $\|h\|_{L^1(B_R)} := \int_{B_R} \|h\| \, d\mathrm{vol}$ and the volume margin are taken, and $\Omega_R := B(0, R)$ in observational space, which holds the clean balls. A *wall* is one of the hyperplanes $\{z : \langle \nu_a, z\rangle = t_a\}$ that cut a PWA map in latent space. The signs of $\langle \nu_a, z\rangle - t_a$ decide which branch domain contains z.

Normalization. We state the result for whitened pairs: p_Z has mean 0 and covariance I. Whitening changes neither the observed law nor (MST) (Proposition 17). It reduces the affine ambiguity of Theorem 8 to a compact gauge: an affine map carrying one whitened law onto another has zero translation and orthogonal linear part, so it lies in $O(d)$.

The class. Fix $\kappa = (N, m, K, M, \pi_0, \lambda_-, \lambda_+, c, \rho, R)$. $\mathcal{F}_\kappa$ is the set of whitened pairs (f, Z) such that (1) p_Z is an irredundant K-component GMM with weights $w_\ell \geq \pi_0$, means $\|\mu_\ell\| \leq M$ and covariances $\lambda_- I \preceq \Sigma_\ell \preceq \lambda_+ I$; (2) f is a reduced PWA map with m branches $F_j(z) = A_j z + b_j$ on domains P_j, cut by at most N walls with unit normals ν_a and offsets $|t_a| \leq M$, with $\sigma(A_j) \subset [1/M, M]$ and $\|b_j\| \leq M$; (3) the four margins below hold.

(AD) with margin c. Atoms have parameters at distance at least c: for $(j, \ell) \neq (j', \ell')$,

$$\|(A_j\mu_\ell + b_j) - (A_{j'}\mu_{\ell'} + b_{j'})\| + \|A_j\Sigma_\ell A_j^\top - A_{j'}\Sigma_{\ell'}A_{j'}^\top\|_F \;\geq\; c.$$

(USB) with margin c. Fix $\vartheta \in (0, 1/2)$ and $\gamma > 1$. A ball $B(x, r) \subseteq \Omega_R$ is a *clean ball* for branch j if (toggle) the branch image $F_j(P_j)$ and its complement each fill at least a fraction ϑ of the volume of $B(x, r)$, and (isolation) every other branch image covers $B(x, \gamma r)$ or misses it, up to null sets. Let β_j be the supremum of the radii of clean balls for j (0 if there are none). The margin is $\beta_j \geq c$ for every j. Toggling keeps the ball from being full, empty or merely tangent, so branch j switches inside it, and isolation says that no other branch does (Lemma 36). Near a boundary both volume fractions tend to $1/2$ as $r \to 0$, which supplies clean balls and caps ϑ at $1/2$. Any such ϑ and γ give the same proof.

(MST) with margin c. $W_1(O_\sharp\, p_Z, p_Z) \geq c$ for every $O \in O(d)$ with $\|O - I\|_{\mathrm{op}} \geq \sqrt{2}$. By the normalization, $O(d)$ is the whole gauge. The cutoff is needed, since $W_1(O_\sharp\, p_Z, p_Z) \to 0$ as $O \to I$. It suffices, since a nontrivial symmetry group of p_Z would contain some O with $\|O - I\|_{\mathrm{op}} \geq \sqrt{2}$ (Lemma 37), so the margin implies (MST), against discrete and continuous symmetries alike.

Volume margin. $\mathrm{vol}(P_j \cap B_R) \geq \rho$ for every branch j.

Proposition 12 is stated for this class. Its ω is a modulus: a nondecreasing function with $\omega(\varepsilon) \to 0$ as $\varepsilon \to 0$, depending only on κ, ϑ and γ, whose existence is asserted without a rate.

The class in coordinates. Fix a map Λ from the 2^N sign patterns of the walls to branch indices, padding with repeated walls so that at most N walls become exactly N. This splits $\mathcal{F}_\kappa$ into finitely many pieces and adds no hypothesis. The two pairs compared may have different Λ. With Λ fixed, a pair is a coefficient vector θ (wall normals and offsets, branch data, GMM data) in a compact product of spheres, closed intervals and closed bounded matrix sets. Write $\theta = (f, Z)$ and $\Phi(\theta) := f_\sharp\, p_Z$.

Lemma 34 (Continuity of the parametrization)**.** *Let $\theta_n \to \theta_\infty$ with Λ fixed. Then (i) $W_1(\Phi(\theta_n), \Phi(\theta_\infty)) \to 0$, (ii) $\|f_n - f_\infty\|_{L^1(B_R)} \to 0$, and (iii) $W_1(p_{Z_n}, p_{Z_\infty}) \to 0$.*

Proof. Converging walls (ν_a, t_a) change half-space volume by $o(1)$ on each ball, and finite Boolean operations preserve symmetric-difference convergence, so $\mathrm{vol}((P_j^{(n)} \triangle P_j^{(\infty)}) \cap B) \to 0$ for every

ball B, where $P_j^{(n)}$ is the domain of branch j at θ_n. For (ii), the affine rules converge uniformly on B_R, and the branch-disagreement set has vanishing measure and bounded integrand. For (iii), write $\mathcal{N}_{\ell,n} := \mathcal{N}(\mu_\ell^{(n)}, \Sigma_\ell^{(n)})$. A common standard normal gives $W_1(\mathcal{N}_{\ell,n}, \mathcal{N}_{\ell,\infty}) \to 0$. Couple matched component mass by these, and move the vanishing weight mismatch at a W_1-cost bounded by the box. For (i), we use the Kantorovich–Rubinstein duality and

$$\int u \, d\Phi(\theta_n) = \sum_{j,\ell} w_\ell^{(n)} \int_{P_j^{(n)}} u\big(F_j^{(n)} z\big) \, d\mathcal{N}_{\ell,n}(z)$$

to reduce the difference to weight, rule, domain and Gaussian-law terms. The first two vanish by coefficient convergence and uniform first moments. Given $\delta > 0$, take S such that the first-moment tails of the Gaussians in the box off $B_S := B(0, S)$ are below δ uniformly in n. On B_S the other two terms vanish by the symmetric-difference convergence above, uniformly bounded Gaussian densities (from $\lambda_- I \preceq \Sigma_\ell$) and Gaussian total-variation convergence. The bounds are uniform over 1-Lipschitz u with $u(0) = 0$, and letting $\delta \downarrow 0$ proves (i). □

Lemma 35 (The (USB) margin is closed)**.** *$\theta \mapsto \min_j \beta_j(\theta)$ is upper semicontinuous, so $\{\theta : \min_j \beta_j(\theta) \geq c\}$ is closed.*

Proof. Let $\theta_n \to \theta_\infty$ with $\min_j \beta_j(\theta_n) \geq c$, fix j, and let $\chi_k^{(n)}$ be the indicator of the branch image $F_k(P_k)$ at θ_n. (1) *A converging clean ball.* The supremum defining β_j need not be attained, so pick at each n a clean ball $B(x_n, s_n) \subseteq \Omega_R$ with $s_n > c - 1/n$, and for each $k \neq j$ a value $v_{k,n} \in \{0, 1\}$ with $\mathrm{vol}(\{\chi_k^{(n)} = v_{k,n}\} \cap B(x_n, \gamma s_n)) = 0$. On a subsequence, $x_n \to x_\infty$, $s_n \to s_\infty \geq c$ and $v_{k,n} = v_k$. Since $\|x_n\| + s_n \leq R$ passes to the limit, $B(x_\infty, s_\infty) \subseteq \Omega_R$. (2) *Volumes pass to the limit.* The walls converge, and $\sigma(A_k) \subseteq [1/M, M]$ makes the inverse branch maps converge. So, off the null set of images of the limiting walls under the limiting branch maps, the signs deciding $\chi_k^{(n)}(x)$ are eventually those of the limit, and $\chi_k^{(n)} \to \chi_k^{(\infty)}$ a.e. Ball indicators converge a.e. when centres and radii converge, so by dominated convergence $\mathrm{vol}(\{\chi_k^{(n)} = v\} \cap B(x_n, r_n)) \to \mathrm{vol}(\{\chi_k^{(\infty)} = v\} \cap B(x_\infty, r))$ whenever $r_n \to r$, for $v \in \{0, 1\}$. (3) *The limit ball is clean.* With $r_n = s_n$, both toggle inequalities pass to the limit. With $r_n = \gamma s_n$, each isolation equality stays 0. So $\beta_j(\theta_\infty) \geq s_\infty \geq c$. □

Lemma 36 (A clean ball gives (USB))**.** *If branch j of f has a clean ball, then there are adjacent chambers of f across which j is the unique entering or leaving branch. In particular, $\beta_j > 0$ for every j implies (USB).*

Proof. Each branch image is a finite union of full-dimensional polyhedra, so its essential boundary lies in finitely many hyperplanes. Let $B(x, r)$ be a clean ball for j. By toggle, $\{\chi_j = 1\}$ and $\{\chi_j = 0\}$ both have positive measure in the ball, which is connected (indecomposable), so the essential boundary of $F_j(P_j)$ contains a relatively open $(d-1)$-dimensional piece of one of those hyperplanes inside the ball, with full-dimensional regions on both sides. By isolation, each χ_k, $k \neq j$, is a.e. constant on $B(x, \gamma r)$. Chambers are open, hence of positive measure, and carry a constant χ_k, so every chamber meeting the ball carries that value. The two chambers across that piece therefore differ exactly in j: they are adjacent, and j is the unique branch entering or leaving. The argument uses chambers, not points, so it does not depend on whether the domains are open, closed or mixed. □

Lemma 37 (No subgroup in the $\sqrt{2}$ ball)**.** *A subgroup $H \leq O(d)$ with $\|O - I\|_{\mathrm{op}} < \sqrt{2}$ for every $O \in H$ is $\{I\}$.*

Proof. In real orthogonal normal form, $\|O - I\|_{\mathrm{op}} = \max_k 2|\sin(\psi_k/2)|$ over the block angles ψ_k, a -1 block counting as $\psi = \pi$. So every angle in the $\sqrt{2}$ ball lies in $[0, \pi/2)$, and any $O \in H$ other than I has an angle $\psi \in (0, \pi/2)$. Take n least with $n\psi \geq \pi/2$. Then $(n-1)\psi < \pi/2$, so $n\psi < \pi/2 + \psi < \pi$, whence $\|O^n - I\|_{\mathrm{op}} \geq 2\sin(n\psi/2) \geq \sqrt{2}$, contradicting $O^n \in H$. □

Proof of Proposition 12. Exact identifiability (Theorem 8, one pair at a time) plus compactness gives one modulus for all pairs.

Step 1: the class is compact. Whitening, the (AD) margin and the volume margin are continuous in θ (the volume by the symmetric-difference convergence of branch domains in the proof of Lemma 34). The (MST) margin is continuous by Berge's maximum theorem, as a minimum over the fixed compact set $\{O : \|O - I\|_{\mathrm{op}} \geq \sqrt{2}\}$ of a jointly continuous function (Lemma 34(iii)). The (USB) margin is closed by Lemma 35. So the coefficient set of each piece of $\mathcal{F}_\kappa$ is closed in a compact set, hence compact.

Step 2: the modulus. For pairs $\theta = (f, Z)$ and $\theta' = (g, Z')$ let

$$\Delta(\theta, \theta') := \min_{O \in O(d)} \max\big(\|f - g \circ O^{-1}\|_{L^1(B_R)},\ W_1(p_Z, O_\sharp\, p_{Z'})\big),$$

and let $\omega(\varepsilon)$ be the supremum of $\Delta(\theta, \theta')$ over $\theta, \theta' \in \mathcal{F}_\kappa$ with $W_1(\Phi(\theta), \Phi(\theta')) \leq \varepsilon$. It is nondecreasing, and finite since the coefficient box bounds both quantities at $O = I$. The minimum is attained: $O(d)$ is compact, the coefficients of $g \circ O^{-1}$ and $O_\sharp\, p_{Z'}$ depend continuously on O and stay in the box, and Lemma 34(ii)–(iii) make the objective continuous. So the proposition holds once $\omega(\varepsilon) \to 0$.

Suppose instead $\omega(\varepsilon) \to \delta_0 > 0$. Then there are $\theta_n = (f_n, Z_n)$ and $\theta'_n = (g_n, Z'_n)$ in $\mathcal{F}_\kappa$ with $W_1(\Phi(\theta_n), \Phi(\theta'_n)) \leq 1/n$ and $\Delta(\theta_n, \theta'_n) \geq \delta_0/2$. (1) *Extract a limit.* There are finitely many Λ, so infinitely many n share one pair of them, and a further subsequence converges, $\theta_n \to \theta_\infty = (f_\infty, Z_\infty)$ and $\theta'_n \to \theta'_\infty = (g_\infty, Z'_\infty)$. By Step 1 both limits lie in $\mathcal{F}_\kappa$. They are still reduced and irredundant, because merged branches or components would give two atoms at parameter distance 0, against the (AD) margin. (2) *Exact equality.* By Lemma 34(i) and the triangle inequality, $W_1(\Phi(\theta_\infty), \Phi(\theta'_\infty)) \leq W_1(\Phi(\theta_\infty), \Phi(\theta_n)) + 1/n + W_1(\Phi(\theta'_n), \Phi(\theta'_\infty)) \to 0$, so the limits satisfy (1). (3) *Exact identifiability.* The margins give (AD) on both sides, and (USB) (Lemma 36) and (MST) for (f_∞, Z_∞). Theorem 8 gives an affine $\bar\alpha$ with $f_\infty = g_\infty \circ \bar\alpha^{-1}$ (Definition 3.3) and $\bar\alpha(Z'_\infty) \sim Z_\infty$. Both laws are whitened, so $\bar\alpha = \bar O \in O(d)$. (4) *Travel back.* Hold $\bar O$ fixed. The coefficients of $g_n \circ \bar O^{-1}$ and of $\bar O_\sharp\, p_{Z'_n}$ converge to those of $g_\infty \circ \bar O^{-1}$ and $\bar O_\sharp\, p_{Z'_\infty}$. By Lemma 34(ii)–(iii) and (3), $\|f_n - g_n \circ \bar O^{-1}\|_{L^1(B_R)} \to 0$ and $W_1(p_{Z_n}, \bar O_\sharp\, p_{Z'_n}) \to 0$. So $\Delta(\theta_n, \theta'_n) < \delta_0/2$ for large n, a contradiction. □

What the margins buy. The proposition changes no theorem. (i) The margin is the conditioning parameter: our conditions are open, and a margin closes the class. The modulus ω is qualitative (compactness gives existence, not a rate), and we claim nothing at margin zero. (ii) On the family of App. S, the margin is a sample complexity. (iii) The margins are computable on a fitted model, returning a number rather than a verdict (§4.5).

S Numerical illustration: full setting

The family has $d = 1$ and its only moving part is the (MST) margin. Let $Z_\eta \sim (\frac{1}{2} + \eta)\mathcal{N}(-1, 1) + (\frac{1}{2} - \eta)\mathcal{N}(1, 1)$, let f have branch slopes $(1, 3, 2)$ and zero intercepts on $(-\infty, -1]$, $[-1, 1]$, $[1, \infty)$, and let the competitor be $h = f \circ S$, with $S(x) = -x$ off $(-1, 1)$ and $S(x) = x$ inside. Measured before whitening, the (AD), (USB) and volume margins do not depend on η, and the (MST) margin is $W_1((-1)_\sharp\, p_{Z_\eta}, p_{Z_\eta}) = 4\eta$. After the whitening of App. R it is 0.110 at $\eta = 0.16$ and 0.0038 at $\eta = 0.005$, also linear in η to leading order. At $\eta = 0$ the two models induce the same law, yet $h \neq f \circ \alpha$ for every affine α. Let KL be the Kullback–Leibler divergence between the two observed laws, and $n^*(\eta) := \chi^2_{1,0.95}/(2\,\mathrm{KL})$ the sample size at which the expected likelihood-ratio statistic $2n\,\mathrm{KL}$ reaches the 5% critical value of χ^2_1. This Wilks calibration is heuristic for two simple models: the exponent below is exact, the constant is not. At n^*, the likelihood-ratio test picks f over h in about 84% of runs (asymptotically $\Phi(\sqrt{\chi^2_{1,0.95}/4}) = 0.8365$; measured in the last row, 400 replicates each, binomial standard error at most 0.020).

η	0.160	0.040	0.010	0.005
Total variation (TV) of the observed laws	0.1527	0.0382	0.0095	0.0048
$n^*(\eta)$	29	460	7 313	29 211
Selection rate at n^*	0.810	0.850	0.840	0.845

Here $\mathrm{TV} = (2\Phi(2) - 1)\eta = 0.9545\eta$ exactly and $\mathrm{KL} \propto \eta^2$ to leading order, so on this family $n^* \propto \eta^{-2}$, the inverse square of the (MST) margin, and power curves collapse when plotted against

$n/n^*(\eta)$. Since n^* is a population quantity, we compute it exactly rather than read it off a training curve. The supplementary material contains the code (`common.py`, `exp_oracle.py`; NumPy only; `python exp_oracle.py`, seed 0).

T Additional Discussion

Boundary richness vs. algebraic conditions. The main-text MID results use (USB) (every branch has an exposed boundary) to establish the branch bijection, illustrating how geometric structure can substitute for algebraic conditions. This is not the only route: Appendix Q presents two alternative theorems that replace (USB) with purely algebraic conditions on the map–latent pair—orbit disjointness ((OD); Theorem 32) and transporter-orbit disjointness ((TOD); Theorem 33). These approaches require weight conditions or stronger orbit conditions instead of boundary richness, illustrating the modularity of the framework.

Globalness, not canonicity, of α. The link α constructed in Theorem 5 comes from a specific toggling pair $(i_\star, j_\star)$ and the permutation σ. Neither the toggling pair nor σ is claimed canonical: LID says only that *some* affine link exists, not that a single canonical one is attached to the truth. What matters for MID is not canonicity but *globalness*—forcing the same structural relationship to hold across *every* branch, not just at the boundary that produced α. The role of (MST) is exactly to deliver this globalness. For a fixed truth–estimator pair, MID produces an affine α with $\alpha(Z') \sim Z$ and $f = g \circ \alpha^{-1}$. Without (MST), different branches could in principle use different affine relabellings of the mixture components; (MST) forces all branchwise difference maps $S_i = f_i^{-1} \circ g_{j(i)}$ into a trivial mixture-symmetry group, so that they all coincide. This is a statement of *global consistency across branches*, not of a "universal" α independent of which estimator one considers. Different estimators matching the same observable law can in general be related to the truth by different affine links.

Necessity of assumptions. For Gaussian mixtures, Appendix O proves that if a GMM admits any nontrivial affine self-symmetry, then it admits a finite-order one (i.e., *torsion*). In globally injective piecewise decoder classes that are closed under precompositions, such torsion symmetries can be localised into law-preserving discontinuous precompositions, producing observationally identical yet mechanism-distinct models. Concretely, a torsion T can give a two-piece construction $S = \{T, \mathrm{Id}\}$ (details omitted here for scope) that yields a non-(AD) strict PWA self-map S of the kind discussed in Remark 4.2, which explains why such S breaks identifiability. This is the mechanism behind the necessity of (MST). We also have complementary results (see Table 5), showing analogous torsion-based necessity phenomena for broad classes of exponential-family and location-scale mixtures, and further examples (omitted here) showing that the same obstruction can persist even in continuous piecewise classes. Boundary richness is likewise necessary: Appendix K.1 gives a simple fold example in which latent (MST) and (AD) both hold, yet (SB) fails and law identifiability breaks.

Towards a broader algebraic identifiability theory. Gaussian mixtures and piecewise-affine maps are the first concrete testbed for a wider program. In the present paper, the relevant ambiguity group is the affine group $\mathrm{Aff}(\mathbb{R}^d)$, which is exactly the natural symmetry group of the Gaussian family. More generally, one can replace affine maps by the natural invertible transformations that preserve a chosen component family F, and then study mixtures from F together with piecewise maps whose branch rules come from that family symmetry group. We also have a group-theoretic proof of (MST) genericity (omitted for scope) that uses weaker conditions than any sufficient conditions shown in Appendix I *and* applies to many exponential-family mixtures. A key next step is to formulate family-level analogues of atom distinctness and to understand how far the branch class can be enlarged before hidden symmetries reappear. Our expectation is that analogues of (AD), (MST), and (USB) will again mark the boundary between identifiability and non-identifiability for a much wider class of piecewise-parametric latent variable models.

U Technical comparisons with related work

This section gathers technical comparisons that are too detailed for the main text.

Two roles of (MST): comparison with prior work. The condition (MST) plays a role analogous to two distinct mechanisms in prior identifiability frameworks: In **iVAE** [25], auxiliary labels (observed class indices) serve as tags that allow local identifiability to extend across the latent space.

In our setting, (MST) achieves a similar effect: the mixture components serve as internal tags, and (MST) ensures their geometry is rich enough to pin down all branch-wise ambiguities. In **Kivva et al.** [26], MID is obtained by a different mechanism: continuity forces adjacent affine pieces to agree on their common boundary, so their relative map fixes a hyperplane. The remaining local alternatives are shear-type or reflection-type. Global density equality, analytic continuation, and GMM identifiability rule out the shear case, since a nontrivial shear cannot preserve a nondegenerate Gaussian mixture. Injectivity then rules out the reflection/fold case. By contrast, our (MST) assumption bypasses this continuity-plus-injectivity argument by directly trivializing the affine self-symmetry group of the latent mixture. Example 2 (Appendix K), adapted from Kivva et al. [26, Example D.2], concretely illustrates how both (AD) and (MST) fail simultaneously, yielding a strict-PWA self-map.

Comparison with iVAE. Under the PWA+GMM model class, the assumptions of iVAE [25] are strictly stronger than our (MST), (USB), and *chamber-wise* (AD): injectivity of the map implies (USB) and chamber-wise (AD) (Appendix M), while observed domain labels together with iVAE's sufficient-variability condition (assumption (iv)) imply (MST) (Appendix N). Concretely: observed labels fix component identities so any admissible symmetry must be component-preserving, reducing (MST) to (JST); the mean/variance variation that (iv) requires then trivialises the remaining within-component stabiliser, giving (JST) and hence (MST). Since our main results use *global* (AD), which is strictly stronger than the chamber-wise version (Appendix E), we cannot say that our full assumption set is strictly weaker than iVAE's. However, our framework strictly weakens the *domain-contrast assumption*: (MST) replaces iVAE's requirement of observed domain labels plus sufficient variability. The weakening is strict. (MST) can hold with as few as $d + 1$ components (affinely independent means, Lemma 23; with labels observed this is already (MST), App. N), whereas (iv) requires $K > 2d$ and does not apply below that. (MST) also needs no labels.

What our conditions trivialize, and what they do not. Were the target group itself trivialized, no ambiguity would remain, i.e., strong identifiability in the sense of Xi and Bloem-Reddy [46], which we do *not* claim. Two groups must be kept apart. In Theorem 4.2 of Squires and Ravikumar [41] the target group is a free parameter. Ours is the admissible choice $\mathrm{Aff}(\mathbb{R}^d)$, and our contrast conditions (domain, mechanism, interaction; Table 1) do not shrink it. *What (MST) makes trivial* is the affine symmetry group of the latent law itself, $G_{\mathrm{mix}}(p)$ (Section 4.3).

Theorem 4.2 of Squires and Ravikumar [41] takes the characterization of the intersection of two transition sets as its *hypothesis*, and for PWA-GMM *that characterization is our whole problem*. We draw the correspondence as framing, not implication. Their valid kernel transitions, in the deterministic case the maps $f^{-1} \circ g$, are what (AD) with (SB)/(USB) constrains, and our conclusion is containment in the target group. Their valid measure transitions reduce to the stabilizer of p, hence to $G_{\mathrm{mix}}(p)$ on its affine part, when the latent law is shared, as in Lemma 7.

The symmetry-breaking view is shared with Squires and Ravikumar [41] and Xi and Bloem-Reddy [46] and is much older [3, 20]. *Our priority* is the conjunction of four features: algebraic symmetry breaking as the engine; the latent law's own symmetry group as the object trivialized; a concrete nonlinear class (PWA decoders, GMM latents); and no supervision. Table 4's proof-flavour column places us against prior work.

(USB) and structural sparsity. A structural-sparsity condition [51] isolates a *source* through the support/intersection pattern of the Jacobian. (USB) isolates a *branch* through a boundary at which exactly one branch toggles. Both are local structural conditions that pin one factor at a time. The difference is the geometry they live in: differential and support-theoretic for smooth decoders, versus boundary-combinatorial for PWA, where the condition survives discontinuity. This also explains why (USB) is strictly weaker than global injectivity: an exposing boundary is needed once per branch, not everywhere.

Boundary richness vs. weak injectivity. Appendix L establishes that (USB) is strictly weaker than global injectivity and logically incomparable with weak injectivity (in the sense of Kivva et al. [26]), with concrete examples in both the continuous and discontinuous regimes.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction state exactly the hierarchy of identifiability results (LID, MID, posterior, pointwise) that are proved as formal theorems in §4 and Appendix A, with conditions precisely matched.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The Limitations paragraph of §5 states the population-level scope, equal latent and observation dimensions, invertible branch rules, conjectural genericity of (SB)/(USB) under continuity, open necessity of (AD), the causal reading as an analogy, and that the stability result has no rate; Necessity and prospects lists what remains open.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All conditions ((AD), (SB), (USB), (MST)) are formally stated in §3–4; proof sketches appear in the main text and complete proofs in Appendix A (Proposition 12: Appendix R). All lemmas used in the proofs are stated and proved within the appendix; statements labelled as conjectures are used by no proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The one numerical illustration (§4.5) is specified in full in App. S: the family, the competitor, the test and the exact formula for the sample size n^*.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplementary material contains the code of the numerical illustration (App. S; NumPy only). No data are used.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: App. S gives every setting of the numerical illustration; no model is trained.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The sample size n^* is computed exactly. The selection rates in App. S are Monte Carlo estimates over 400 replicates; their binomial standard error, at most 0.020, is stated there.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The numerical illustration (App. S) is a small NumPy computation that runs on a laptop CPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is foundational mathematical research on identifiability theory; it raises no ethical concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational theory with no direct path to specific applications or deployments; there is no immediate positive or negative societal impact beyond advancing the mathematical foundations of trustworthy generative modelling.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No data, models, or other assets with misuse risk are released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The only external asset is NumPy (BSD-3-Clause license), used by the code of the numerical illustration (Appendix S); all cited work is referenced via standard academic citation.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The only new asset is the code of the numerical illustration (`common.py`, `exp_oracle.py`) in the supplementary material, documented in Appendix S and in the scripts' docstrings (NumPy only; `python exp_oracle.py`, seed 0). No datasets or models are released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper involves no human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used for writing, editing, and as a verification aid during proof development.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.