On the Identifiability of Quantized Factors

Vitória Barin-Pacela
FAIR, Meta; Mila; DIRO, Université de Montréal

Kartik Ahuja
FAIR, Meta

Simon Lacoste-Julien
Mila; DIRO, Université de Montréal; Canada CIFAR AI Chair

Pascal Vincent
FAIR, Meta; Mila; DIRO, Université de Montréal; CIFAR

Editors: Francesco Locatello and Vanessa Didelez

Abstract

Disentanglement aims to recover meaningful latent ground-truth factors from the observed distribution solely, and is formalized through the theory of identifiability. The identifiability of independent latent factors has been proven to be impossible in the unsupervised i.i.d. setting under a general nonlinear map from factors to observations. In this work, however, we demonstrate that it is possible to recover *quantized* latent factors under a generic nonlinear diffeomorphism. We only assume that the latent factors have *independent discontinuities* in their density, without requiring the factors to be statistically independent. We introduce this novel form of identifiability, termed *quantized factor identifiability*, and provide a comprehensive proof of the recovery of the quantized factors.

Keywords: identifiability, disentanglement, causal representation learning, quantized representations, discrete representations

1. Introduction

A large part of intelligence is based on the ability to make sense of observed sensory data without explicit supervision. The goal of representation learning is, thus, to detect and model relevant structure in the distribution of observed data, and expose it into useful compact representations, to facilitate generalization and sample-efficient learning of subsequent tasks. One long-standing goal in that respect has been that of structuring the representation into *disentangled factors* (Bengio et al., 2013). These may be conceived of as "natural" ground truth, descriptive, or causal variables that underlie the observations. A vector representation consisting of recovered disentangled factors may be viewed as corresponding to a natural Cartesian coordinate system for the observations, whereby each varying factor is associated with an axis.

Identifiability theory formalizes the foundations of disentanglement by precisely delimiting the conditions under which it is possible. Unsupervised disentanglement of latent factors has been found impossible in the general nonlinear setting in the absence of further inductive bias (Locatello et al., 2019). This result echoes an older identification impossibility result on nonlinear Independent Component Analysis (Hyvärinen and Pajunen, 1999). As a result, much subsequent work has sidestepped the issue either via stronger inductive biases, such as more restrictive assumptions on the function that maps latent factors to observations (Buchholz et al., 2022; Kivva et al., 2022; Ahuja et al., 2022c; Brady et al., 2023; Lachapelle et al., 2023), for instance sparsity of its Jacobian (Moran

et al., 2022; Zheng et al., 2022; Zheng and Zhang, 2023), or by turning to weakly supervised disentanglement, using some form of additional information (see related works in Appendix A).

Provided that they corresponds to valid assumptions, inductive biases should undoubtedly be used in practice whenever available, as well as any additional supervisory signals. However, in the present theoretical work, we revisit and tackle the problem of fully unsupervised identifiability of latent factors, the most challenging setting. We assume a generic smooth invertible nonlinear mapping: a diffeomorphism. No additional assumptions are made on the mapping, and the assumption of the factors being mutually independent is also discarded.

Given the previous theoretical impossibility results for unsupervised identifiability under a diffeomorphism, a shift in our approach was necessary. We relax the notion of identifiability of continuous factors to that of identifiability of quantized continuous factors.

The promise of quantized, grid-like representations has been argued empirically in both machine learning and neuroscience. It has been suggested that the brain of humans and other animals organizes spatial knowledge and relational concepts into codes that have an hexagonal grid-like pattern (Constantinescu et al., 2016; Whittington et al., 2020). In representation learning, vector quantization has shown enormous success in image generation (van den Oord et al., 2017). Concurrent studies investigate empirically this explicit relationship with disentanglement (Hsu et al., 2023), and further explore quantization in grid-structured representations (Mentzer et al., 2024; Irie et al., 2023; Friede et al., 2023).

However, none of these provide a supporting identifiability theory for quantized factors. In the present work, we first formalize this novel relaxed form of identifiability. We, then, provide a full proof of the identifiability of quantized factors under a general diffeomorphism. This is achieved by assuming, rather than the mutual independence of factors, the presence of *independent discontinuities* in the joint probability density [1] of the latent factors.

Our contributions are the following:

- We introduce and formalize a novel relaxed form of representation identifiability: *quantized factor identifiability*.
- We provide the first proof of representation identifiability under a general diffeomorphic map, which sets itself apart from the impossibility results that dominate the field.

We hope that this novel theoretical foundation may provide useful insights to develop algorithms of practical relevance for robustly learning disentangled representations.

2. From precise factor identifiability to quantized factor identifiability

In this section, we define and contrast the standard form of factor identifiability, which we term "precise factor identifiability", with our new relaxed "quantized factor identifiability" paradigm.

2.1. Setup

We suppose that we have access to observations in $\mathcal{X} \subset \mathbb{R}^D$. They are realizations of the vector random variable $X = (X_1, \dots, X_D)$, which is assumed to be a transformation of a real vector of unobserved latent factors $Z = (Z_1, \dots, Z_d)$, i.e. $X = f(Z)$, via a *bijective* mapping $f : \mathcal{Z} \to \mathcal{X}$ where $\mathcal{Z} \subset \mathbb{R}^d$. The mapping f is called the *mixing map*, which is unknown but is assumed to belong to a broad function class. The latent factors Z follow a distribution represented by the probability

1. More precisely, these are *non-removable discontinuities* in the PDF, as will be elaborated in Section 5.1.

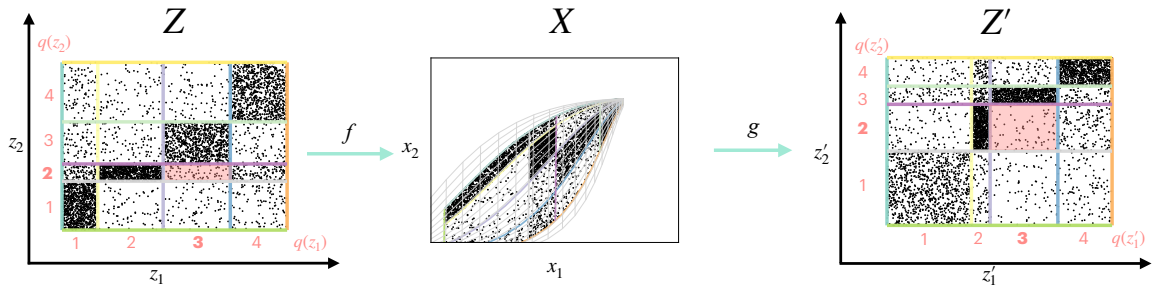


Figure 1: **Recovery of quantized factors. Left:** The true (continuous) latent factors Z_1 and Z_2 are not independent, but their joint probability density p_Z has *independent discontinuities*: sharp changes in the density that are aligned with the axes and form a grid. **Middle:** The factors get warped and entangled by the diffeomorphism f into observations X, but the discontinuities in their density survive in the observed space. **Right:** We can learn a diffeomorphism g that yields a density $p_{Z'}$ having axis-aligned discontinuities. This suffices to recover a grid whose cells match the initial grid's cells (up to possible permutation and axis reversal). Pink cell **example:** the points Z' in cell $(3, 2)$ originated from the points Z in cell $(3, 2)$. To construct these cells, the quantization of each continuous factor to an integer depends on thresholds based on the location of the discontinuities. The quantizations of Z_1' and Z_2' match precisely the quantizations of Z_1 and Z_2, up to possible permutation and axis reversal. This summarizes the *identifiability of quantized factors* under diffeomorphisms.

density function (PDF) p_Z, which is also unknown but is typically subject to assumptions. This induces a distribution for X whose PDF is denoted by p_X. The mapping $g : \mathcal{X} \to \mathcal{Z}$ approximates f^{-1} at the optimum.

The setup is summarized in the following diagram.

$$\underbrace{Z}_{\substack{\subset \mathbb{R}^d \\ \text{true factors}}} \sim p_Z \xrightarrow[\text{unknown}]{f} \underbrace{X}_{\substack{\subset \mathbb{R}^D \\ \text{observed data}}} \xrightarrow[\text{learned}]{g \approx f^{-1}} \underbrace{Z'}_{\substack{\subset \mathbb{R}^d \\ \text{recovered factors}}} \qquad h = g \circ f$$

In identifiability theory, the distribution of observations p_X is supposedly known. Alternatively, for this level of precision, observed samples from p_X can be considered with the sample size approaching infinity. Identifiability theorems also need to make clear assumptions on the mixing map f and on the density of factors p_X.

In the remainder of this section, we first formalize the usual factor identifiability as well as our proposed relaxation to quantized factor identifiability in the general case. Subsequently, we focus on the case where f is a diffeomorphism.

2.2. Precise Factor Identifiability

The usual, precise factor identifiability theorems amount to statements of the following form:

Precise Identifiability of Factors: Knowledge of p_X is sufficient to determine a reverse mapping $g : \mathbb{R}^D \to \mathbb{R}^d$ that will yield recovered factors $(Z'_1, \ldots, Z'_d) = g(X)$ that correspond one-to-one to the ground-truth factors $(Z_1, \ldots, Z_d)$, up to permutation and component-wise invertible transformations (ideally monotonic).

Formally: there exists an indices permutation function σ and invertible scalar functions γ_i such that $\forall i \in \{1, \ldots, d\}, \gamma_i(Z'_i) = Z_j$ with $j = \sigma(i)$. Precise factor identifiability theorems require specifying assumptions on f and on p_Z.

2.3. Quantization of factors

Let us now specify how the factors can be quantized. For simplicity, we consider that each factor is a real-valued scalar. A real number z may be quantized to an integer based on a tuple of real thresholds T via the following quantization operation:

$$Q(z; T) = \sum_{k=1}^{|T|} \mathbf{1}_{z \geq T_k}. \tag{1}$$

For example, consider $z \in [0, 4]$ and $T = (0.5, 2.0)$. Then $Q(z; T) = \mathbf{1}_{z \geq 0.5} + \mathbf{1}_{z \geq 2.0}$. So

$$Q(z; T) = \begin{cases} 0, & 0 \leq z < 0.5 \\ 1, & 0.5 \leq z < 2 \\ 2, & 2 \leq z \leq 4. \end{cases}$$

We also define quantization with order reversal as $Q^-(z; T) = \sum_{k=1}^{|T|} \mathbf{1}_{z \leq T_k}$. For convenience, we will use the notation $Q^{(s)}$ to mean Q if $s = +1$ and Q^- if $s = -1$.

The set of specific thresholds used for quantizing a random variable Z_i is typically derived from some properties of its distribution p_{Z_i} (e.g. a set of $|T|$ specific quantiles). We will consider the more general case, where the thresholds for quantizing Z_i might be determined not only based on p_{Z_i}, but more generally, on i and on the joint probability density of all factors p_Z. The operation returning a set of thresholds to be used for a factor Z_i is denoted by $\mathcal{T}(p_Z, i)$. Thus, the quantization of Z_i may be written as: $q_i(Z_i) = Q(Z_i; \mathcal{T}(p_Z, i))$.

2.4. Quantized Factor Identifiability

Quantized factor identifiability theorems will be statements of the following form:

Identifiability of Quantized Factors: Knowledge of p_X is sufficient to determine a reverse mapping $g : \mathbb{R}^D \to \mathbb{R}^d$ that will yield recovered factors $(Z'_1, \ldots, Z'_d) = g(X)$ such that their quantization $(q'_1(Z'_1), \ldots, q'_d(Z'_d))$ will correspond one-to-one to the quantized ground-truth factors $(q_1(Z_1), \ldots, q_d(Z_d))$, up to possible permutation of indices and order reversal.

Formally, there exists an indices permutation function σ and order-reversal indicators $s_i \in \{-1, +1\}$ such that: $\forall i \in \{1, \ldots, d\}$, $q'_i(Z'_i) = q_j(Z_j)$, with $j = \sigma(i)$, where q_j and q'_i are monotonic quantization functions. We can, more precisely, define q_j as $q_j(Z_j) = Q(Z_j; \mathcal{T}(p_Z, j))$, and $q'_i(Z'_i) = Q^{s_i}(Z'_i; \mathcal{T}(p_{Z'}, i))$. The precise operation $\mathcal{T}$ that determines how the quantization

thresholds are obtained from properties of the distributions remain to be specified by the particular quantized factor identifiability theorem.

Hence, quantized factor identifiability theorems require specifying assumptions on f, assumptions on p_Z, as well as a precise quantization operation. For the quantization to be meaningful, it should produce at least two non-empty bins. That is, for any given factor, the respective factor samples will be mapped to at least two different quantized values. Quantization to a single all-encompassing bin is trivially identifiable and useless.

We highlight that quantized factor identifiability does not intend to prove identifiability when the true factors take a discrete set of values. Instead, we define a relaxed form of identifiability for *continuous* ground-truth factors. Quantization leads to a loss of precision/resolution, resulting in a coarser identification.

3. What to assume on p_Z when f is a diffeomorphism

From now on, we will turn our attention to the case where the mixing map f is assumed to be a general *diffeomorphism*, that is, a continuously differentiable function with a continuously differentiable inverse. The goal is to learn the approximate inverse diffeomorphism g. First, let us discuss what assumptions we should make on the distribution of factors p_Z that may yield a positive identifiability result.

3.1. Disentanglement, independence, and discontinuities

Disentanglement has been equated to finding statistically independent factors (Khemakhem et al., 2020a), but statistical independence has been criticized as an unrealistic and problematic assumption (Träuble et al., 2021; Dittadi et al., 2021; Roth et al., 2023) whose association to disentanglement is misleading. For example, the usual *descriptive factors* with which we describe scenes are usually not statistically independent. Consider the variables color, shape, and background: bananas tend to be yellow; cows tend to be on grass backgrounds; camels tend to be on sand.

Moreover, a primary interest for learning disentangled representations is as an enabler of robust generalization under distribution shifts. From this perspective, we should aim for factor discovery approaches that are stable and insensitive to broad changes in the (unknown) distribution of the factors, such as whether they happen to be independent or correlated in the data. Aiming for extracting statistically independent factors will, by construction, be very sensitive to this, which goes contrary to the desired robustness.

Lastly, and most importantly for our goal of characterizing what form of unsupervised identification is possible under a diffeomorphism, assuming statistical independence is insufficient, as previous impossibility results for nonlinear ICA have shown (Hyvärinen and Pajunen, 1999). This is fundamentally due to the extreme flexibility of diffeomorphisms. Even more discouraging, Buchholz et al. (2022) have shown that even if we knew p_Z precisely, we could not achieve precise factor identifiability. This is because a diffeomorphism can move data points along isosurfaces of p_Z while keeping the same p_Z, thus rendering entangled representations indistinguishable from disentangled ones.

To prevent this movement along isolines of p_Z from taking points from one region to another of the factor space, there could be barriers of discontinuity separating different regions of p_Z. We develop this justification for the need for discontinuities more precisely in Appendix G, based on the result from Buchholz et al. (2022).

Another perspective to consider is that discontinuities are among the few characteristics of a density that diffeomorphisms can neither erase nor create (Theorem 4). Hence, they are good candidates for holding cues in p_Z guaranteed to survive in some form when mapped to X via any diffeomorphism. Thus, if they were indicative of coordinate axes in Z, there is a prospect of recovering them from the resulting discontinuities in p_X.

Therefore, to enable the identifiability of quantized factors under a diffeomorphism, we will not assume that p_Z implies statistically independent factors, but rather that it has *independent discontinuities*, which we define precisely in the next section.

3.2. Independent discontinuities in the probability density

Here, we contrast the statistical independence of factors with our approach, the independence of discontinuities. We will assume that there are discontinuities in the PDF of the factors, and that the location of these discontinuities in the density of any given factor is independent of the values of all the other factors.

Definition 1 *Let $\mathcal{S}$ be the support of p_Z. We say that p_Z has an **independent discontinuity** at $Z_i = \tau$ when every point in the intersection of the* coordinate hyperplane $\{\mathbf{z}_i = \tau\}$ *with $\mathcal{S}$ is a non-removable discontinuity of p_Z. Formally, this independent discontinuity at $Z_i = \tau$ is defined as the set $\Gamma_{\mathcal{S}}(i, \tau) = \{\mathbf{z} \in \mathcal{S} | \mathbf{z}_i = \tau\}$ under the condition that $\forall \mathbf{z} \in \Gamma_{\mathcal{S}}(i, \tau)$, p_Z has a non-removable discontinuity at $\mathbf{z}$.*

Such discontinuities are "independent" in the sense that we have a discontinuity at $Z_i = \tau$ regardless of the values taken by the other factors. Only the locations of the discontinuities in the density need to be independent of the other factors. This does not impose statistical independence of the factors, nor anything else wherever the density is continuous. Thus, assuming the presence of independent discontinuities can accommodate statistically independent factors as well as correlated factors.

Geometrically, an independent discontinuity in p_Z corresponds to a *coordinate hyperplane* restricted to the support of p_Z. This hyperplane is orthogonal to the Z_i axis and parallel to all the other axes. We will, thus, interchangeably call it an **independent discontinuity** or **axis-aligned discontinuity**.

For our theorems, we will further require that the interior of the support of the density is connected. A connected independent discontinuity that splits this support in two is said to be an **axis-separator** (formally defined in Section 5.2). If the set of all non-removable discontinuities of p_Z is the union of a finite set of such axis-separators, with at least one along each axis, then we say that they form an **axis-aligned grid**. Figure 1 (left) gives an example of two factors that are clearly not statistically independent but that have independent discontinuities in their PDF, appearing to the eye as axis-aligned discontinuities along each axis. Altogether, they from an axis-aligned grid.

Independent discontinuities are striking landmarks in the PDF landscape p_Z that remain detectable in p_X and $p_{Z'}$. A diffeomorphic map, even though it can warp the space in almost arbitrary ways, will not be able to erase such discontinuities. These are the robust cues that we can rely on to achieve quantized factor disentanglement under a diffeomorphism.

4. Overview of the main quantized identifiability result

In a nutshell:

Assumptions

- f is a diffeomorphism
- $(Z_1, \ldots, Z_d) \sim p_Z$ are d continuous random variables.
- The interior of the support of p_Z is a connected set.
- The set of non-removable discontinuities of p_Z is the union of a finite set of independent discontinuities – at least one along each dimension – that together form an *axis-aligned grid*. This grid must also possess a *backbone* (precisely defined in the next section).

Quantized factor identifiability theorem Under the above assumptions:

- It suffices to learn a diffeomorphism g yielding $Z' = g(X)$ such that the PDF of $p_{Z'}$ has independent discontinuities forming an axis-aligned grid.
- Then, the quantized reconstructed factors $(q'_1(Z'_1), \ldots, q'_d(Z'_d))$ will correspond one-to-one to the quantized ground-truth factors $(q_1(Z_1), \ldots, q_d(Z_d))$, up to possible permutation of indices (and order reversal).
- The quantization thresholds used for q_i and q'_i are obtained as the locations of the independent discontinuities.

This result is illustrated in Figure 1. The formal **Quantized factor identifiability theorem** is Theorem 17, found in section 5.3 together with its proof. It builds on two other theorems: the *Non-removable discontinuity preservation theorem* (Theorem 4 in Section 5.1) and the *Grid structure recovery theorem* (Theorem 15 in Section 5.2) and its corollary. The following section presents these theorems and the required definitions in their logical order.

5. Main theorems

Most of the theory will concern the diffeomorphism $h := g \circ f$ that maps Z to Z'.

5.1. Non-removable discontinuity preservation theorem

We will show that discontinuities in the PDF are preserved by a a diffeomorphism. However, one subtlety is that the PDF corresponding to a given distribution is not unique, as elaborated in Appendix F. The many PDFs representing the same distribution actually form an equivalence class, whose elements may take arbitrarily different values on sets of points of measure zero. So not all the discontinuities in a PDF are meaningful. Since we care about observable characteristics of the actual distribution, we must focus on aspects of the PDF that are immune to erasure by changes of measure zero. We use the following definitions:

Definition 2 ***Removable discontinuity:*** *A PDF p has a removable discontinuity at z if p is discontinuous at z but there exists another p' in the same equivalence class (i.e. p and p' yield the exact same probability measure) that is continuous at z.*

Definition 3 ***Non-removable discontinuity:*** *A PDF p has a non-removable discontinuity at z if p is discontinuous at z but this discontinuity is not removable. Equivalently,* all *PDFs in the equivalence class of p are discontinuous at z. Note that a non-removable discontinuitiy is a property of an equivalence class of PDFs, thus of the distribution, not just of a single PDF.*

Theorem 4 ***Non-removable discontinuity preservation theorem.*** *Let Z be a latent random variable with values in $\mathcal{Z} \subset \mathbb{R}^d$, whose distribution is represented by a PDF p_Z. Let $h : \mathcal{Z} \to \mathcal{Z}' \subset \mathbb{R}^d$ be a diffeomorphism, and let $Z' = h(Z)$ be a transformed random variable whose distribution is represented by a probability density function $p_{Z'}$. Then, $p_{Z'}$ has a non-removable discontinuity at a point z' if and only if p_Z has a non-removable discontinuity at the point $z = h^{-1}(z')$.*

Proof: Appendix E.1.

5.2. Grid structure recovery theorem and corollary

5.2.1. DEFINITION OF GRID STRUCTURE

The notions we use to define the grid structure are related to usual hyperplanes and hypersurfaces of $\mathbb{R}^d$, but they are restricted to a connected subset $\mathcal{S}$ of $\mathbb{R}^d$. In our setting, $\mathcal{S}$ will be the interior of the support of the density on which the grids can be defined. Let $\mathcal{S} \subset \mathbb{R}^d$ be a connected open smooth submanifold of dimension d ($\mathcal{S}$ such as an open d-ball). We will use the following definitions, which are **illustrated in Appendix D**.

Definition 5 *The **splitting** of a set $\mathcal{S}$ by another set $\mathcal{C}$, denoted* $\mathrm{split}(\mathcal{S}, \mathcal{C})$*, is the set of* connected components *of $\mathcal{S} \setminus \mathcal{C}$.*

Definition 6 *We say that $\mathcal{C}$ **splits** $\mathcal{S}$ **in two** to mean $|\mathrm{split}(\mathcal{S}, \mathcal{C})| = 2$ (we denote the cardinality of a countable set A by $|A|$, and similarly for the number of elements in an ordered list or a tuple).*

Definition 7 *We say that $\mathcal{C}$ is a **separator** of $\mathcal{S}$ if $\mathcal{C}$ is a* connected subset *of $\mathcal{S}$ and $\mathcal{C}$ splits $\mathcal{S}$ in two. The two connected components that result from the split are called the two **halves** resulting form the split, denoted $\mathcal{C}^+$ and $\mathcal{C}^-$, i.e. $\{\mathcal{C}^+, \mathcal{C}^-\} = \mathrm{split}(\mathcal{S}, \mathcal{C})$.*

Definition 8 *$\mathcal{C}$ is a **smooth separator** of $\mathcal{S}$ if $\mathcal{C}$ is a* separator *of $\mathcal{S}$ and is a* smooth hypersurface *of $\mathcal{S}$ (i.e. a smooth embedded submanifold of dimension $d - 1$).*

Definition 9 *An **axis-separator** of $\mathcal{S}$ is a special case of* smooth separator *of $\mathcal{S}$ that is the intersection of $\mathcal{S}$ with an* axis-aligned hyperplane *of $\mathbb{R}^d$ (a coordinate hyperplane). It can be defined as $\mathcal{H} = \Gamma_{\mathcal{S}}(i, \tau) = \{z \in \mathcal{S} | z_i = \tau\}$ (Figure 7). Because it is a separator, it splits $\mathcal{S}$ in two halves $\Gamma^+_{\mathcal{S}}(i, \tau) = \{z \in \mathcal{S} | z_i > \tau\}$ and $\Gamma^-_{\mathcal{S}}(i, \tau) = \{z \in \mathcal{S} | z_i < \tau\}$, which are each nonempty and connected* (Figure 8).

Definition 10 *An **axis-separator-set** $\mathcal{G}$ on $\mathcal{S}$ is a finite set of axis separators of $\mathcal{S}$.*

Definition 11 *An axis-aligned **grid** $G \subset \mathcal{S}$ is a subset of $\mathcal{S}$ that can be obtained as a union of all the separators in an axis-separator-set $\mathcal{G}$. i.e. $G = \cup\mathcal{G} = \cup_{H \in \mathcal{G}} H$.*
Note the important distinction we make between a *grid*, which is a subset of $\mathcal{S}$ and hence a set of points, and an *axis-separator-set*, which is a set of axis separators (which themselves are sets of

points). An *axis-separator-set* thus has more explicit structure than a $grid$. The proof we will unroll depends conceptually on the ability to rebuild, in several steps, the entire grid internal structure, starting from only the unstructured *grid* as a set of points. The first step of this program will be the recoverability of *axis-separator-set* from *grid.*

Definition 12 *A **parallel-separator-set** is a* set *of axis-separators all defined on the same i^{th} axis (and are thus parallel). In particular, we denote the subset of axis-separator set $\mathcal{G}$ that are all defined on the i^{th} axis as $\mathcal{G}^{(i)}$.*

Definition 13 *A **discrete coordination** $\mathbf{A}$ is a tuple $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_d)$ where each $\mathbf{A}_i$ is itself a tuple of real numbers in increasing order $\mathbf{A}_i = (\mathbf{A}_{i,1}, \ldots, \mathbf{A}_{i,n_i})$ such that $\mathbf{A}_{i,k+1} > \mathbf{A}_{i,k}$. These represent the coordinates of axis-separators along each of the d coordinate axes (Figure 10).*
Note: $\mathbf{A}_i$ contains the list of quantization *thresholds* to quantize the i^{th} coordinate (or factor) as $Q(Z_i; \mathbf{A}_i)$, as defined in equation 1.
A discrete coordination defines the entire grid structure. One can easily obtain the various constituent sets from it:

(a) the individual *separators* (≈"hyperplanes") $\Gamma_{\mathcal{S}}(i, \mathbf{A}_{i,k})$, and their positive and negative halves (≈"half-spaces") $\Gamma^{+}_{\mathcal{S}}(i, \mathbf{A}_{i,k})$ and $\Gamma^{-}_{\mathcal{S}}(i, \mathbf{A}_{i,k})$ respectively;
(b) the *parallel-separator-sets* $\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(d)}$, where $\mathcal{G}^{(i)} = \{\Gamma_{\mathcal{S}}(i, \mathbf{A}_{i,k})\}_{k=1}^{|\mathbf{A}_i|}$;
(c) the *axis-separator-set* $\mathcal{G} = \mathcal{G}^{(1)} \cup \ldots \cup \mathcal{G}^{(d)}$;
(d) the *grid* $G = \mathrm{grid}_{\mathcal{S}}(\mathbf{A}) = \cup\mathcal{G}$.

Definition 14 *A **backbone** $\mathcal{H}^*$ of a grid is a* list $\mathcal{H}^* = (\mathcal{H}^*_1, \ldots, \mathcal{H}^*_d)$ *of d separators of that grid, each defined on the corresponding axis, that have a non-empty intersection (they meet at a single point z^*). That is, for $\mathcal{H}^*_1 \in \mathcal{G}^{(1)}, \ldots, \mathcal{H}^*_d \in \mathcal{G}^{(d)}$, we must have $\bigcap_{i=1}^{d} \mathcal{H}^*_i = \{z^*\}$. In addition, for $\mathcal{H}^*$ to be a backbone, it is also required that each of its separators $\mathcal{H}^*_i$ intersect* all *the other separators $H \in \mathcal{G}^{(j)}$ of the grid that are defined on the other axes $j \neq i$ (those not in the same parallel-separator-set); namely, $\forall i, \forall j \neq i, \forall H \in \mathcal{G}^{(j)}, \mathcal{H}^*_i \cap H \neq \emptyset$* (example in Figure 11).
A backbone functions as a set of "main axes", and we will require a proper **grid** to have at least one backbone. This is a weaker requirement than requiring a "complete grid" where *each separator of the grid* would be required to intersect all the separators that are not in the same parallel-separator-set. Here, we require only that the separators of the backbone intersect all the other separators on the other axes of the grid.

5.2.2. GRID STRUCTURE PRESERVATION AND RECOVERY THEOREM

Theorem 15 ***Grid structure preservation and recovery theorem.*** *Let $h : \mathcal{S} \subset \mathbb{R}^d \to \mathcal{S}' \subset \mathbb{R}^d$ be a diffeomorphism, where both $\mathcal{S}$ and $\mathcal{S}'$ are open connected subsets of $\mathbb{R}^d$. Suppose we have an axis-aligned grid $G \subset \mathcal{S}$, associated with its axis-separator-set $\mathcal{G}$ and discrete coordination $\mathbf{A}$, that is, $G = \mathrm{grid}_{\mathcal{S}}(\mathbf{A})$. While the grid does not need to be "complete", we suppose that $\mathcal{G}$ has at least one* backbone*. Now, suppose that we have another axis-aligned grid in $\mathcal{S}'$, associated with its discrete coordination $\mathbf{B}$, with $G' = \mathrm{grid}_{\mathcal{S}'}(\mathbf{B})$. Suppose $G' = h(G)$. Then, there exists a permutation function σ over dimension indexes $1, \ldots, d$ and a direction reversal vector $s \in \{-1, +1\}^d$ such that $\forall j \in \{1, \ldots, d\},\ i = \sigma^{-1}(j),\ K = |\mathbf{A}_i| = |\mathbf{B}_j|, \forall k \in \{1, \ldots, K\}, \forall z' \in \mathcal{S}'$,*

If $s_i = +1$, then:

$$\begin{cases} z'_j = \mathbf{B}_{j,k} \iff h^{-1}(z')_i = \mathbf{A}_{i,k}, \\ z'_j > \mathbf{B}_{j,k} \iff h^{-1}(z')_i > \mathbf{A}_{i,k}, \\ z'_j < \mathbf{B}_{j,k} \iff h^{-1}(z')_i < \mathbf{A}_{i,k}; \end{cases}$$

If $s_i = -1$, then:

$$\begin{cases} z'_j = \mathbf{B}_{j,k} \iff h^{-1}(z')_i = \mathbf{A}_{i,K-k+1}, \\ z'_j > \mathbf{B}_{j,k} \iff h^{-1}(z')_i < \mathbf{A}_{i,K-k+1}, \\ z'_j < \mathbf{B}_{j,k} \iff h^{-1}(z')_i > \mathbf{A}_{i,K-k+1}. \end{cases}$$

Principle of the proof Starting from the premise $G' = h(G)$, we know that h maps every point of G to a point of G'. The proof recovers the entire underlying grid *structure* in 3 steps:

Step 1 recover a one-to-one mapping of the individual separators: $\mathcal{G}' = h(\mathcal{G})$.

Step 2 recover the partition into subsets of parallel separators (each subset associated to an axis): $\mathcal{G}'^{(j)} = h(\mathcal{G}^{(i)})$ (with permutation $j = \sigma(i)$).

Step 3 show that the ordering of the separators in a parallel-separators-set is preserved (up to possible order reversal): $[h(\Gamma_{\mathcal{S}}(i, \mathbf{A}_{i,1})), \dots, h(\Gamma_{\mathcal{S}}(i, \mathbf{A}_{i,K}))] = [\Gamma_{\mathcal{S}'}(j, \mathbf{B}_{j,1}), \dots, \Gamma_{\mathcal{S}'}(j, \mathbf{B}_{j,K})]$ or in reversed order $[h(\Gamma_{\mathcal{S}}(i, \mathbf{A}_{i,1})), \dots, h(\Gamma_{\mathcal{S}}(i, \mathbf{A}_{i,K}))] = [\Gamma_{\mathcal{S}'}(j, \mathbf{A}_{j,K}), \dots, \Gamma_{\mathcal{S}'}(j, \mathbf{A}_{j,1})]$. And similarly, the ordering of the halves corresponding to each of these separators is preserved. Knowing to which half (either $\Gamma^+_{\mathcal{S}'}(j, \tau)$ or $\Gamma^-_{\mathcal{S}'}(j, \tau)$) a point $\mathbf{z}'$ belongs tells us whether z'_j is above or below the threshold τ.

For example, seeing that $z'_j > \mathbf{B}_{j,k}$ tells us that $z' \in \Gamma^+_{\mathcal{S}'}(j, \mathbf{B}_{j,k})$, which implies from step 3 (in the case of no order reversal) that its preimage $z = h^{-1}(z')$ belongs to $\Gamma^+_{\mathcal{S}}(i, \mathbf{A}_{i,k})$, which yields $z_i > \mathbf{A}_{i,k}$. This is what Theorem 15 expresses. We refer the reader to Appendix E.3 for the full proof.

Corollary 16 ***Recovery of quantized factors.*** *Under the same premises as Theorem 15, consider random variables Z and $Z' = h(Z)$. Using the quantization operation Q (previously defined in Section 2.3, equation 1), we recover quantized factors up to permutation σ of the axes and possible direction reversal indicated by s: $\forall i \in 1, \dots, d$, $Q(Z_i; \mathbf{A}_i) = Q^{s_i}(Z'_j; \mathbf{B}_j)$ with $j = \sigma(i)$.*

Proof $Z' = h(Z)$ implies that $Z_i = h^{-1}(Z')_i$.
Now if $s_i = +1$, Theorem 15 yields $Z'_j \geq \mathbf{B}_{j,k} \iff Z_i \geq \mathbf{A}_{i,k}$.
Thus, $Q(Z_i, \mathbf{A}_i) = \sum_k \mathbf{1}_{Z_i \geq \mathbf{A}_{i,k}} = \sum_k \mathbf{1}_{Z'_j \geq \mathbf{B}_{j,k}} = Q(Z'_j; \mathbf{B}_j)$.
Similarly, if $s_i = -1$, Theorem 15 yields $Z'_j \leq \mathbf{B}_{j,k} \iff Z_i \geq \mathbf{A}_{i,k}$.
Thus, $Q(Z_i, \mathbf{A}_i) = \sum_k \mathbf{1}_{Z_i \geq \mathbf{A}_{i,k}} = \sum_k \mathbf{1}_{Z'_j \leq \mathbf{B}_{j,k}} = Q^-(Z'_j; \mathbf{B}_j)$.
So in both cases, we have $Q(Z_i; \mathbf{A}_i) = Q^{s_i}(Z'_j; \mathbf{B}_j)$. ■

5.3. Quantized factor identifiability theorem

Theorem 17 ***Quantized factors identifiability theorem.*** *Let Z be a latent random variable with values in $\mathcal{Z} \subset \mathbb{R}^d$ and whose PDF is p_Z. Let $f : \mathcal{Z} \to \mathcal{X} \subset \mathbb{R}^D$ be a diffeomorphism, and $X = f(Z)$*

be the observed random variable. Assume that the support of the PDF p_Z is an open connected set[2]*. Further assume that p_Z has at least one connected independent discontinuity in each dimension, such that the set of non-removable discontinuities of p_Z forms an axis-aligned grid with a backbone. Let $\mathbf{A}$ be the discrete coordination of this grid. Then, there exists a diffeomorphism $g : \mathcal{X} \to \mathcal{Z}'$ yielding a variable $Z' = g(X)$ such that the set of non-removable discontinuities of the PDF $p_{Z'}$ is an axis-aligned grid. Consider any such diffeomorphism g, and let $\mathbf{B}$ be the discrete coordination of its resulting axis-aligned grid. Then, there exists a permutation function σ over the dimension indexes $1, \ldots, d$, and a direction reversal vector $s \in \{-1, +1\}^d$ such that $q'_j(Z'_j) = q_i(Z_i)$ with $i = \sigma^{-1}(j)$, where $q'_j(Z'_j) = Q^{s_i}(Z'_j; \mathbf{B}_j)$ and $q_i(Z_i) = Q(Z_i; \mathbf{A}_i)$. In other words, the quantized factors in Z' agree with the quantized factors in Z, up to permutation and possible axis reversal.*

Proof Note that existence is trivial (it suffices to take $g = f^{-1}$, which yields $Z' = Z$). But the fact that *any* g that yields a PDF whose non-removable discontinuities form an axis-aligned grid will have this property can now easily be proven from our previous results. It suffices to consider $h = g \circ f$ to be a diffeomorphism (the composition of two diffeomorphisms), so that $Z' = h(Z)$, and to combine the non-removable discontinuity preservation theorem (Thm. 4) with the grid structure preservation and recovery theorem (Thm. 15). Let $G = \text{grid}_{\mathcal{S}}(\mathbf{A})$ and $G' = \text{grid}_{\mathcal{S}}(\mathbf{B})$ be the set of non-removable discontinuity points of p_Z and $p_{Z'}$, respectively. From the non-removable discontinuity preservation theorem, we have that $G' = h(G)$. And from the grid structure preservation and recovery theorem and its corollary, we have that $G' = h(G)$ implies that there exists a permutation function σ over dimension indexes $1, \ldots, d$ and a direction reversal vector $s \in \{-1, +1\}^d$ such that $Q^{s_i}(Z'_j; \mathbf{B}_j) = Q(Z_i; \mathbf{A}_i)$ with $i = \sigma^{-1}(j)$. We have, thus, proved that the quantized factors of Z' agree with the quantized factors of Z, up to permutation and axis reversal. ■

6. Independent discontinuities in real-world disentangled factors

We have motivated independent discontinuities as a theoretical requirement to be able to identify quantized factors even after they passed through highly flexible diffeomorphic maps. In real data under finite samples, we can only hope for a smooothed density estimate that will never show true discontinuities, but merely sharp changes (gradients of large magnitude) in the density. Also if one is willing to assume a slightly less flexible map, such as Lipschitz, the requirement for true discontinuities may likely be relaxed to merely sharp changes.

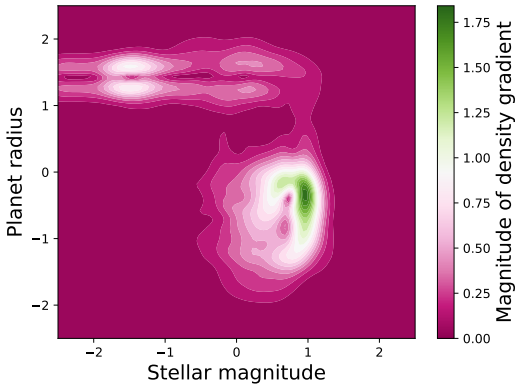


Figure 2: Grid structure observed in the PDF of the NASA exoplanet dataset (standardized log of factors).

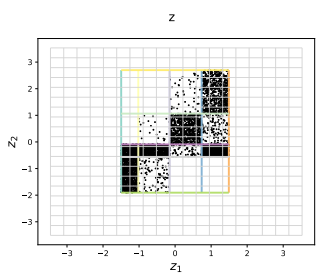
Still, one may wonder why and how such sharp density changes could appear in latent factors of real-world data. Here is a simple example: due to gravity, people and most objects tend to be either in a standing or lying position. One will seldom see them with a $45°$ pitch angle

2. Alternatively, if the support is not open, we can consider its interior.

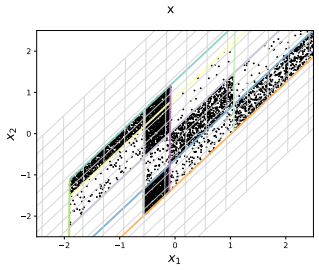
irrespective of how other factors appear (e.g. background color). This results in a sharp change (discontinuity) in the PDF of the pitch angle factor, independent of the values of the other factors. This is an example of a density jump due to a physical equilibrium point, of which we can expect many variants in nature.

Empirically, we found evidence of independent sharp density changes forming a grid structure in descriptive factors of the NASA Exoplanet Archive (Akeson et al., 2013). Figure 2 shows the magnitude of the gradient of the density of the factors *stellar magnitude* and *planet radius*. Locations of high magnitude gradient show an axis-aligned grid, compatible with independent jumps in the density similar to the synthetic data from Figure 1. We provide another evidence of axis-alignment in real motion-capture data in Appendix C.

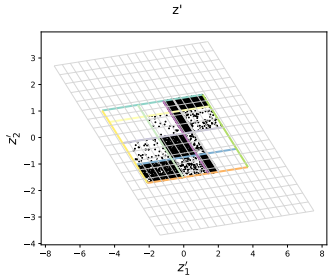
7. Experiments



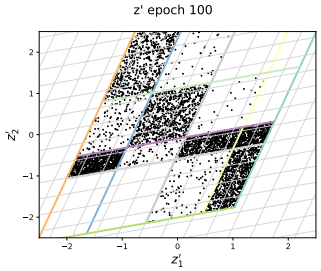
(*a*) True factors.



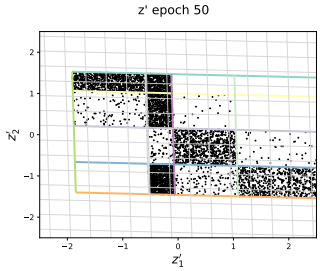
(*b*) Observed variables.



(*c*) Linear ICA reconstruction of the factors.



(*d*) Reconstruction of the factors by Hausdorff Factorized Support.



(*e*) Our model's reconstruction of the factors.

Figure 3: The true latent factors (a) **do not have factorized support** and are **correlated**. The observations (b) are the result of a **linear** map applied to the factors. Our method (c) obtains a factorized representation corresponding to the ground-truth factors. Both linear ICA (d) and Hausdorff Factorized Support (e) Roth et al. (2023) fail to learn the axis-aligned true latent factors.

We develop a criterion for learning axis-aligned discontinuities and present a proof-of-concept experiment for the case where the mixing map f is linear. This is a simple prototype to demonstrate the feasibility of quantized identification based on independent discontinuities in the PDF, and how it can be advantageous compared to other methods already in the linear case. We present a tentative

criterion for nonlinear transformations in Appendix B.2, but it remains to be thoroughly tested and experimented. We reserve the proposal and analysis of a full practical criterion for future work.

The method we present here aims to align the gradients of the joint density of the factors with the axes. We perform a density estimation $\hat{p}_\sigma$ of Z and use it to obtain the gradients $\frac{\partial \log \hat{p}_\sigma}{\partial z}$. Gradients of high magnitude hint at potential discontinuities. We encourage the alignment of these gradient vectors with the standard basis vectors (axes) by maximizing their cosine similarity. The algorithm and experimental setup is detailed in Appendix B.1.

Figure 3 presents the visualization of the reconstruction of the latent variables for our model, compared to Linear ICA (using the FastICA algorithm (Hyvärinen, 1999)) and to Hausdorff Factorized Support (HFS) (Roth et al., 2023), for the case where the ground-truth latent factors are neither independent nor have a factorized support. The results show that the true latent structure is well-reconstructed by our model. The learned factor grid is axis-aligned and corresponds to the original grid up to permutation and axis reversal, as anticipated by the theory. The quantized cells are correctly identified up to this indeterminacy. Meanwhile, both FastICA and Hausdorff Factorized Support learn the factors up to a rotation and shearing (besides permutation and scaling), because their reconstruction is not axis-aligned. Appendix B.1 presents the results in the case where the support is factorized and we remark that our model is again able to axis-align the factors, while even HFS' reconstruction does not present factorized support [3].

8. Conclusion and future work

In this theoretical work, we have introduced the novel paradigm of *quantized factor identifiability*. We have then shown that *fully unsupervised identifiability* of quantized factors is possible under *diffeomorphisms*. This is significant given that the prevailing literature is dominated by impossibility results. We are able to achieve the identification of quantized factors, provided that we assume independent discontinuities in the latent factor's distribution (which naturally form a grid). The novel relaxed (weaker) form of identifiability is meant as a step towards more realistic assumptions for disentanglement: no restrictive inductive bias on the mapping and no assumed independence of factors, rather aiming for potential causal footprints (Lopez-Paz et al., 2017).

However, there are important limitations to this theory, the most obvious being that it requires actual *discontinuities*. This is required due to the flexibility of general diffeomorphisms, as justified in Appendix G. Future work shall try to relax this to just sharp (but not infinitely sharp) changes in the density, under slightly less general Lipschitz smooth mappings.

When moving to the finite sample setting, we must resort to density estimation, which yields a smoothed estimate of p_X, and as a result, discontinuities will become non-infinite sharp changes. These "softer" discontinuities can still be detected by considering the magnitude of the gradient of the density (as we display in Figure 2). The development of an effective practical training criterion and algorithm to train a nonlinear reverse mapping g to recover an axis-aligned grid is left for future work (Appendix B.2 proposes a possible starting direction).

3. The code for reproducing these results is available at https://github.com/facebookresearch/quantized_identifiability/.

Acknowledgments

The authors thank Léon Bottou for sharing his original motivation for discontinuities in the density from a causal perspective, as well as David Lopez-Paz for related discussions on causal footprints. The present work only barely mentions these motivations due to its identifiability focus, but we are grateful for Léon's and David's encouragement in exploring this direction. The authors thank Diane Bouchacourt for discussions on Theorem 1 and the experimental validation of the theory on real-world datasets, as well as Mark Ibrahim for his contribution to early discussions while this project was taking shape. We also thank Sébastien Lachapelle for feedback on this project, and Mohammad Pezeshki for feedback on the paper.

Vitória Barin-Pacela is partially supported by the Canada CIFAR AI Chair Program, as well as a grant from Samsung Electronics Co., Ldt., administered by Mila, in support of her PhD studies at the University of Montreal. Simon Lacoste-Julien and Pascal Vincent are CIFAR Associate Fellows in the Learning in Machines & Brains program.

This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program.

References

Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. Weakly Supervised Representation Learning with Sparse Perturbations. In *Advances in Neural Information Processing Systems*, 2022a.

Kartik Ahuja, Divyat Mahajan, Vasilis Syrgkanis, and Ioannis Mitliagkas. Towards efficient representation identification in supervised learning. In *1st Conference on Causal Learning and Reasoning*, 2022b.

Kartik Ahuja, Yixin Wang, Divyat Mahajan, and Yoshua Bengio. Interventional causal representation learning. In *40th International Conference on Machine Learning*, 2022c.

R. L. Akeson, X. Chen, D. Ciardi, M. Crane, J. Good, M. Harbut, E. Jackson, S. R. Kane, A. C. Laity, S. Leifer, M. Lynn, D. L. McElroy, M. Papin, P. Plavchan, S. V. Ramí rez, R. Rey, K. von Braun, M. Wittman, M. Abajian, B. Ali, C. Beichman, A. Beekley, G. B. Berriman, S. Berukoff, G. Bryden, B. Chan, S. Groom, C. Lau, A. N. Payne, M. Regelson, M. Saucedo, M. Schmitz, J. Stauffer, P. Wyatt, and A. Zhang. The NASA exoplanet archive: Data and tools for exoplanet research. *Publications of the Astronomical Society of the Pacific*, 125(930):989–999, 2013.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Jack Brady, Roland S. Zimmermann, Yash Sharma, Bernhard Schölkopf, and Wieland and von Kügelgen, Julius Brendel. Provably Learning Object-Centric Representations. In *International Conference on Machine Learning*, 2023.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, 2022.

Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function Classes for Identifiable Nonlinear Independent Component Analysis. In *Conference on Neural Information Processing Systems*, 2022.

Marek Capinski and Peter Ekkehard Kopp. *Measure, Integral and Probability*. Springer Undergraduate Mathematics Series. Springer London, 2013. ISBN 9781447106456.

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

Alexandra O. Constantinescu, Jill X. O'Reilly, and Timothy E. J. Behrens. Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468, 2016.

Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wuthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. On the Transfer of Disentangled Representations in Realistic Settings. In *International Conference on Learning Representations*, 2021.

Manfredo P. do Carmo. *Differential Geometry of Curves and Surfaces*. Prentice-Hall, 1976.

David Friede, Christian Reimers, Heiner Stuckenschmidt, and Mathias Niepert. Learning disentangled discrete representations. *arXiV preprint arXiv:2307.14151*, 2023.

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.

Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvarinen. Disentangling Identifiable Features from Noisy Data with Structured Nonlinear ICA. In *Advances in Neural Information Processing Systems*, 2021.

Kyle Hsu, Will Dorrell, James C. R. Whittington, Jiajun Wu, and Chelsea Finn. Disentanglement via Latent Quantization. In *Neural Information Processing Systems*, 2023.

Antti Hyttinen, Vitória Barin-Pacela, and Aapo Hyvärinen. Binary independent component analysis: a non-stationarity-based approach. In *38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

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

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, 2017.

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

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Kazuki Irie, Róbert Csordás, and Jürgen Schmidhuber. Topological Neural Discrete Representation Learning à la Kohonen. In *ICML 2023 Workshop: Sampling and Optimization in Discrete Space*, 2023.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, 2020a.

Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ICA. *Advances in Neural Information Processing Systems*, 2020b.

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

Daniel A. Klain and Gian-Carlo Rota. *Introduction to Geometric Probability*. Cambridge University Press, 1997.

David A. Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021.

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

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

Lek-Heng Lim, Ken Sze-Wai Wong, and Ke Ye. The Grassmannian of affine subspaces. *Foundations of Computational Mathematics*, 21:537—-574, 2021.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, 2022.

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.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, 2020.

David Lopez-Paz, Robert Nishihara, Soumith Chintalah, Bernhard Schölkopf, and Léon Bottou. Discovering Causal Signals in Images. In *Computer Vision and Pattern Recognition*, 2017.

Jerrold E. Marsden and Michael J. Hoffman. *Elementary Classical Analysis*. W. H. Freeman, 1993.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite Scalar Quantization: VQ-VAE Made Simple. In *International Conference on Learning Representations*, 2024.

Gemma E. Moran, Dhanya Sridhar, Yixin Wang, and David M. Blei. Identifiable Deep Generative Models via Sparse Decoding. *Transactions on Machine Learning Research*, 2022.

Karsten Roth, Mark Ibrahim, Zeynep Akata, Pascal Vincent, and Diane Bouchacourt. Disentanglement of Correlated Factors via Hausdorff Factorized Support. In *International Conference on Learning Representations*, 2023.

A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, 47(10):2807–2820, 1999.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017.

Yixin Wang and Michael I Jordan. Desiderata for representation learning: A causal perspective. *arXiv preprint arXiv:2109.03795*, 2021.

James CR Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The tolman-eichenbaum machine: Unifying space and relational memory through generalisation in the hippocampal formation. *Cell*, 183(5), 2020.

Weiran Yao, Guangyi Chen, and Kun Zhang. Learning latent causal dynamics. *arXiv preprint arXiv:2202.04828*, 2022.

Yujia Zheng and Kun Zhang. Generalizing Nonlinear ICA Beyond Structural Sparsity. In *37th Conference on Neural Information Processing Systems*, 2023.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the Identifiability of Nonlinear ICA: Sparsity and Beyond. In *36th Conference on Neural Information Processing Systems*, 2022.

APPENDIX

Appendix A. Related work

We categorize the existing literature on causal representation learning into the following two categories: i) the theory imposes assumptions on both the mixing map and the latent factors, leading to typically fully unsupervised models; ii) the theory imposes assumptions on the distribution of latent factors and not strictly on the mixing map, leading to models that mostly require weak supervision or auxiliary variables. None of these studies considered the recovery of quantized factors like we do in this work.

Identifiability of latent factors in the unsupervised i.i.d setting: In linear Independent Component Analysis (ICA), Comon (1994) established that under a linear and invertible mixing map and independent non-Gaussian latent factors, these latent factors ca be identified up to order and scale indeterminacies. Beyond the linear case, Taleb and Jutten (1999) analyze a post-nonlinear mapping, obtaining the same indeterminacies as in linear mixtures. Gresele et al. (2021) demonstrated that with independent latent factors and a mixing function that adheres to the independent mechanism assumption, some of the non-identifiability counterexamples highlighted in Hyvärinen and Pajunen (1999) can be avoided. Expanding on the role of mixing maps, Buchholz et al. (2022) scrutinized different classes of maps that restrict the Jacobian of the mixing maps. Their study specifically focused on conformal maps and orthogonal coordinate transformations. Kivva et al. (2022) proposes that when the mixing map is piece-wise linear and the latent distribution is a Gaussian mixture (with latent components conditionally independent given a discrete, unobserved confounder), the true latent factors can be identified up to scaling and permutation as well. Ahuja et al. (2022c) asserted that the true latent factors can be identified, barring permutation and scaling errors, when the mixing map is polynomial and latent factors satisfy the support independence assumption, as proposed in Wang and Jordan (2021); Roth et al. (2023). Brady et al. (2023); Lachapelle et al. (2023) obtain identifiability for additive decoders, while Moran et al. (2022); Zheng et al. (2022); Zheng and Zhang (2023) obtain identifiability by assuming a sparse structure on the Jacobian of the mixing function.

Identifiability of latent factors with weak supervision: Research in this category largely makes assumptions on the latent distribution but imposes few constraints on the mixing map. To compensate for this lack of restrictions, these studies necessitate additional information, typically in one of two forms: a) identification driven by auxiliary information (e.g., labels, time stamps), or b) identification driven by weak supervision (e.g., data augmentations) (Hyvarinen and Morioka, 2017; Hyvärinen et al., 2019; Hyvarinen and Morioka, 2016). A key example of auxiliary information-driven identification is the work on identifiable variational autoencoders (Khemakhem et al., 2020a), which assumes the existence of an additionally observed variable such that the latent variables are conditionally independent given it, and the conditional probability density of the latent variable given this auxiliary variable comes from an exponential family. This work has been expanded upon in several subsequent studies (Khemakhem et al., 2020b; Lachapelle et al., 2022; Ahuja et al., 2022b; Hyttinen et al., 2022), which modify some of its assumptions. Locatello et al. (2020); Klindt et al. (2021) assumed access to paired data, which can emerge from data augmentation or natural video frames with sparse changes, resulting in supervision-driven identification. Several follow-up studies (Hälvä et al., 2021; Ahuja et al., 2022a; Brehmer et al., 2022; Yao et al., 2022; Lippe et al., 2022) have built upon this work, moving beyond the independence assumptions on the latent factors and incorporating general transition dynamics.

Appendix B. Practical criterion

B.1. Proof-of-concept experimentation for linear maps

Synthetic data generation:

We generate a grid of points by establishing a prior for each cell, such that the sum of the priors of all the cells equals 1. We define a 4×4 grid, the position of each separator being drawn uniformly inside the range of the grid. In order to generate correlated data, first we draw the prior probabilities from a standard Uniform distribution. Then, we redefine the prior probability of the cells in the diagonal to be higher than the probability of the other cells, followed by normalization. The dataset is composed of 50,000 samples from this distribution. In the dataset with unfactorized support (Figure 3), the correlation coefficient between the true factors of variation is of 0.61.

Algorithm:

The steps for implementing the training criterion are:

1. Randomly initialize a parametric mapping $g : \mathcal{X} \to \mathcal{Z}$ to be learned.
2. From the matrix of observed samples $\mathbf{X}$ of size $n \times D$, compute the matrix of estimated latent variables $\mathbf{Z} = g(\mathbf{X})$ of size $n \times d$ – where g is applied separately to each row[4].
3. Estimate the density of $\mathbf{Z}$ using a kernel density estimator (Parzen window) $\hat{p}_\sigma$ and compute the gradient $\mathbf{V}_{i,\cdot} = \frac{\partial \log \hat{p}_\sigma}{\partial z}(\mathbf{Z}_{i,\cdot})$ at every point of $\mathbf{Z}$.
4. Define importance weighting terms α based on the gradient magnitudes $\alpha_i = \frac{\|V_{i,\cdot}\|}{\sum_{i'=1}^{n} \|V_{i',\cdot}\|}$. A large magnitude of the gradient indicates a sharp jump, hence, this weight indicates how close the sample is to a density jump, that is, how likely it belongs to an axis-separator of the grid.

Let $\bar{\mathbf{V}}_i = \frac{\mathbf{V}_i}{\|\mathbf{V}_i\|}$ be the normalized version of $\mathbf{V}_i$. For the individual gradient vectors to be axis-aligned, the maximum cosine similarity with the canonical axis vectors should be maximized:

$$\text{maximize} \max_{j \in \{1,\ldots,d\}} |\text{cosim}(V_{i\cdot}, \mathbf{1}_j)| = \max_j \frac{|\mathbf{V}_{ij}|}{\|\mathbf{V}_{i\cdot}\|_2} = \frac{\|\mathbf{V}_{i\cdot}\|_\infty}{\|\mathbf{V}_i\|_2} = \|\bar{\mathbf{V}}_{i\cdot}\|_\infty. \tag{2}$$

Then, the final loss function to be optimized over all the points is:

$$\text{minimize } \ell_{\text{grad}-\text{axis}} = -\sum_{i=1}^{n} \alpha_i \|\bar{\mathbf{V}}_{i,\cdot}\|_\infty. \tag{3}$$

Details of model and algorithm – Dataset with unfactorized support:

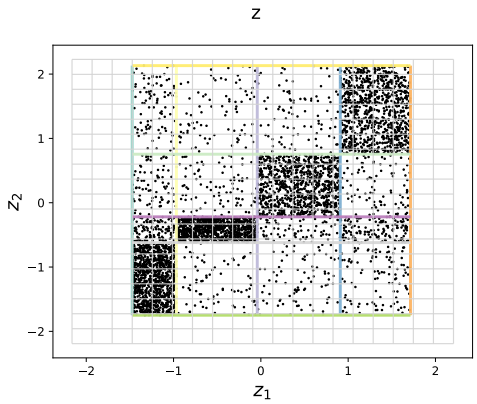
We minimize this loss using stochastic gradient descent with a learning rate of 0.1, momentum of 0.9, and a batch size of 5000 samples. Mini-batches are employed due to the high memory cost of loading the full dataset. The results displayed are for when the training loss stops decreasing.

The kernel density estimation employs a bandwidth of 0.1. With finite samples, we use a density estimator $\hat{p}_{Z'}$, since we do not have access to the exact $p_{Z'}$. We remark that any density estimation will result in some smoothing of the true distribution. So even if there were real discontinuities in the exact density, they will appear as smoothed discontinuities: the gradients of the density have large magnitude, not infinite magnitude.

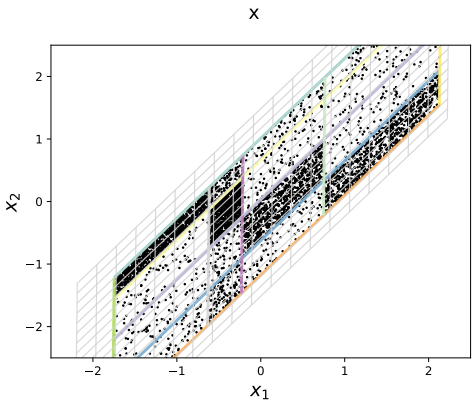
Hausdorff Factorized Support training details:

We train HFS using the `hausdorff_hard` distance approximation which is used throughout the experiments from Roth et al. (2023). In this simple linear case, we simply optimize to minimize

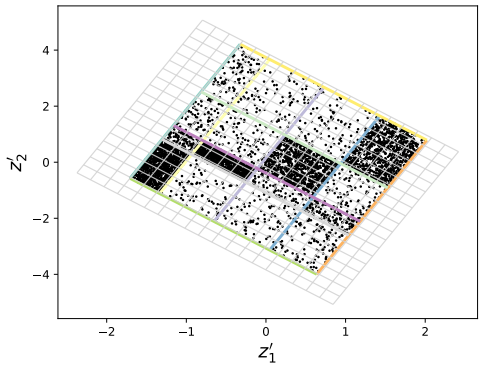
4. Note that we dropped the apostrophe ' in $\mathbf{Z}$ to lighten notation.



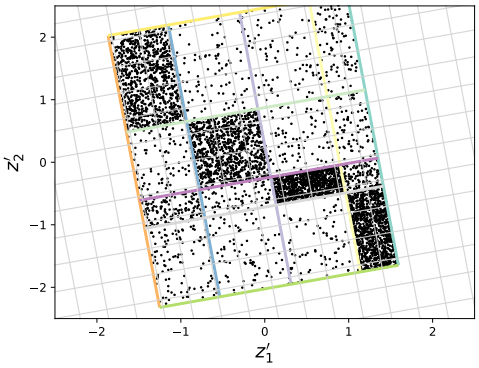
(*a*) True factors.



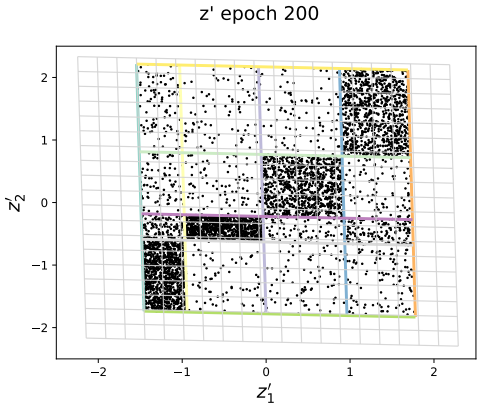
(*b*) Observed variables.



(*c*) Linear ICA reconstruction of the factors.



(*d*) Reconstruction of the factors by Hausdorff Factorized Support.



(*e*) Our model's reconstruction of the factors.

Figure 4: When the true latent factors (4(*a*)) are correlated, our method (4(*e*)) obtains a factorized representation corresponding to the ground-truth factors, as opposed to linear ICA (4(*c*)) and Hausdorff Factorized Support (4(*d*)) which reconstruct the factors up to a rotation.

the Hausdorff distance between the learned factors and their counterpart with factorized support. We did not find an advantage in using the reconstruction term as the representation does not collapse into a single point. Training is done using stochastic gradient descent with a step size of 0.0001 and a batch size of 5000 samples.

Experiment with factorized support: We conduct a similar study on a dataset in which the support of the true factors is factorized. In this dataset (Figure 4, the true factors have a correlation coefficient of 0.64. In this experiment, we attempt to have a fair comparison with HFS and demonstrate that using (discontinuity) information from inside the support can help achieve better axis-alignment (and as a result, better factorized support) of the learned factors of variation.

We compare our model with linear ICA and show that our model is able to learn a factorized representation of the factors, while Fast ICA (Hyvärinen, 1999) fails due to the correlation of the factors violating the independence assumption, as illustrated in Figure 4. HFS also learns the reconstructed factors up to a rotation (but no shearing), even though its factorized support assumption is satisfied, showing that in this case our criterion is effective in aligning the factors with the axes.

B.2. Towards a criterion for nonlinear maps

Alignment of discontinuities in the joint density For nonlinear maps, only encouraging the gradients to be axis aligned does not suffice because the distortions yield a curved latent space. It is also desirable to straighten this deformed grid, which can then be axis-aligned. Here, we outline a few terms that could encourage this behavior in the training dynamics. We can align both the point samples and their gradient vectors. Moreover, the alignment comes in two forms: local alignment in a neighborhood of points, and alignment to the axes.

- **Gradient local alignment term:** encourage pairs of neighboring points of high gradient magnitude to have gradients aligned by maximizing their cosine similarity. We can make the criterion a weighted average of cosine similarities, with significant weights only if they are neighboring points and both have large gradient magnitudes):

$$\begin{aligned}
\beta_{i,i'} &= \alpha_i \alpha_{i'} \exp\left(-\frac{1}{2\sigma_2^2}\|\mathbf{Z}_i - \mathbf{Z}_{i'}\|^2\right) \\
\bar{\beta}_{i,i'} &= \frac{\beta_{i,i'}}{\sum_{i,i'} \beta_{i,i'}} \\
\text{maximize} &\sum_{i,i'} \bar{\beta}_{i,i'} \operatorname{cosim}(\mathbf{V}_i, \mathbf{V}_{i'}) \\
i.e.\ \text{minimize } \ell_{\text{grad}-\text{local}} &= -\sum_{i,i'} \bar{\beta}_{i,i'} \left\langle \bar{\mathbf{V}}_i, \bar{\mathbf{V}}_{i'} \right\rangle
\end{aligned}$$

- **Points local axis alignment term:** encourages neighboring points with large density gradient magnitude to lie on or close to the same axis separator. For this, it suffices that they share one of their coordinates. In other words, is suffices to minimize the minimum over coordinates of the squared difference:

$$\text{minimize } \ell_{\text{points}-\text{local}} = \sum_{i,i'} \bar{\beta}_{i,i'} \min_j \left(\frac{\mathbf{Z}_{ij} - \mathbf{Z}_{i'j}}{\|\mathbf{Z}_i - \mathbf{Z}_{i'}\|} \right)^2$$

- **Points-gradient-orthogonality term:** encourages the gradient vector to be orthogonal to the vectors joining neighboring points by penalizing their squared cosine similarity:

$$\text{minimize } \ell_{\text{points-grad}} = \sum_{i,i'} \bar{\beta}_{i,i'} \left(\left\langle \bar{\mathbf{V}}_i, \frac{\mathbf{Z}_{i'} - \mathbf{Z}_i}{\|\mathbf{Z}_{i'} - \mathbf{Z}_i\|} \right\rangle \right)^2 .$$

We can, then, define a training criterion that is a weighted sum of these terms (with appropriate sign), possibly together with the minimization of a reconstruction error ℓ_{rec} (from a decoder network $\hat{f}$ that tries to reconstruct $\mathbf{X}$ from $\mathbf{Z}$).

$$\ell_{\text{rec}} = \frac{1}{n} \sum_{i=1}^{n} \|\hat{f}(\mathbf{Z}_i) - \mathbf{X}_i\|^2$$

The complete loss to minimize is, thus,

$$L(\theta) = \lambda_1 \ell_{\text{grad-local}} + \lambda_2 \ell_{\text{grad-axis}} + \lambda_3 \ell_{\text{points-local}} + \lambda_4 \ell_{\text{points-grad}} + \lambda_5 \ell_{\text{rec}}$$

where θ is the set of (network) parameters of both encoder g and decoder $\hat{f}$.

Appendix C. Additional evidence of axis-aligned discontinuities in real data

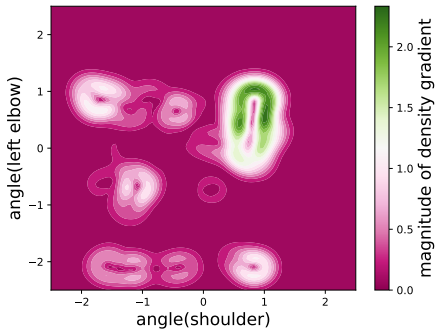
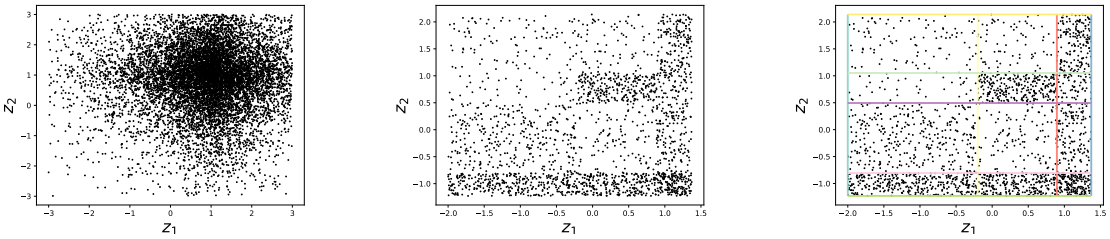


Figure 5: Evidence of axis-aligned discontinuities found in the CMU Motion Capture Dataset.

Figure 5 presents additional evidence of axis-aligned discontinuities in real data from the CMU Motion Capture dataset (obtained from `mocap.cs.cmu.edu`). We preprocess the variables to obtain the angles of the joints of the body in different frames captured. In particular, the angle of the left elbow is the angle formed by the markers "LELB", "LUPA", and "LWRA". The variable angle(shoulders) measures the angle of the shoulders (defined by the markers "RSHO" and "LSHO") with respect to the vertical axis. Then these angles are standardized. In the plot, we can observe axis-aligned discontinuities in green, which represents a high magnitude of the gradient of the density.

Appendix D. Illustrations of the definitions

Figures 7, 8, 10, 9, and 11 illustrate the concepts used in the definitions of section 5.2.



(*a*) Independent non-Gaussian factors

(*b*) Latent factors whose PDF has axis-aligned discontinuities

(*c*) Underlying grid structure

Figure 6: Illustration of different kinds of assumptions on the distribution of latent factors. **Left:** samples from traditional assumption of independent non-Gaussian factors (here using a truncated Laplace distribution). **Middle:** samples from a distribution that follows our assumption of axis-aligned discontinuities in the probability density. **Right:** Underlying grid structure revealing discontinuities in the density landscape as colored *axis-separators*, forming a *grid*. Traditional independence assumption yields non-identifiability result under general nonlinear smooth mapping (diffeomorphism). Our assumption yields, under diffeomorphism, provable recovery of a discretized coordinate system. It allows to map back observed points into the proper latent grid cell – a novel relaxed form of identifiability, which we term *quantized factor identifiability*.

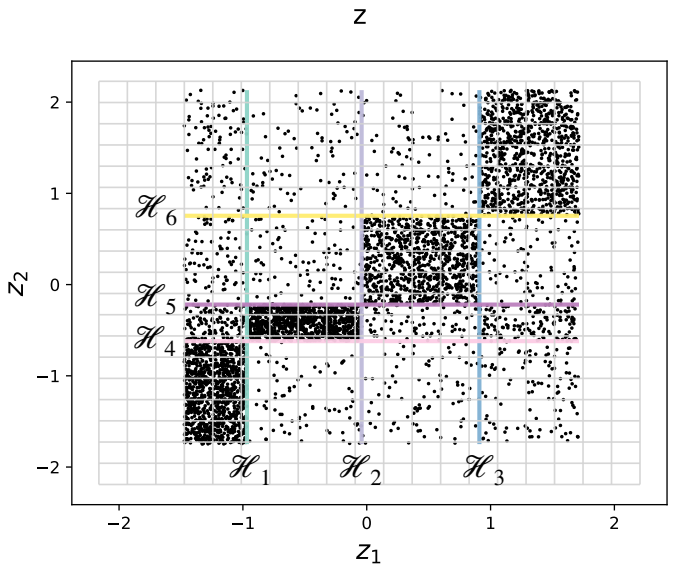


Figure 7: Axis separators $\mathcal{H}_1, \ldots, \mathcal{H}_6$ of $\mathcal{S}$ (Definition 9).

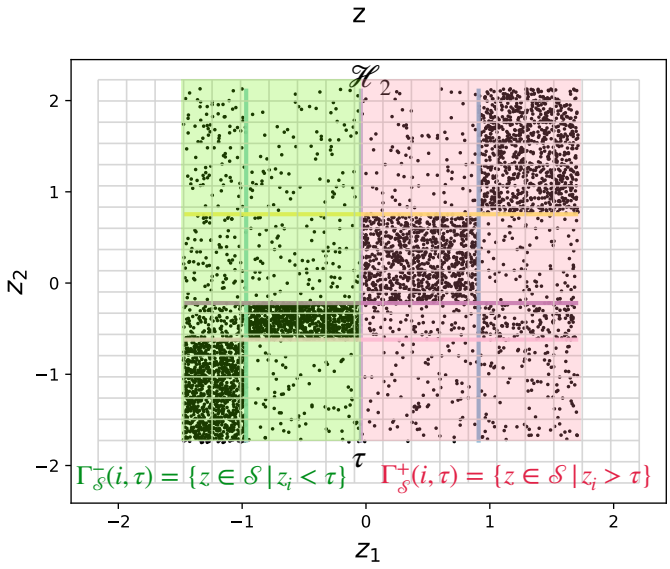


Figure 8: An **axis-separator** of $\mathcal{S}$ splits $\mathcal{S}$ in two halves $\Gamma^+_{\mathcal{S}}(i,\tau) = \{z \in \mathcal{S} | z_i > \tau\}$ and $\Gamma^-_{\mathcal{S}}(i,\tau) = \{z \in \mathcal{S} | z_i < \tau\}$, which are each nonempty and connected (Definition 9).

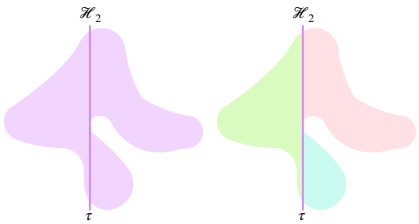


Figure 9: Axis-separator of $\mathcal{S}$ counterexample: Even though the support of the set on the left is connected, $\mathcal{H}_2$ splits it into three parts, not two halves, so it does not satisfy the axis-separator condition (Definition 9).

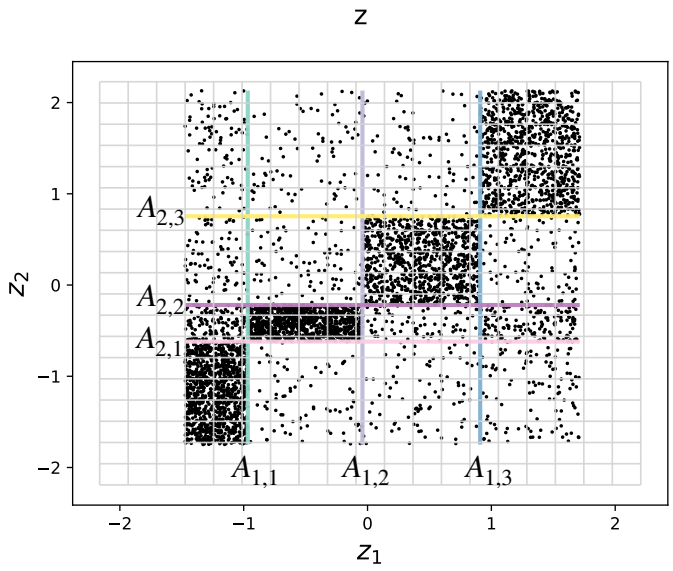


Figure 10: A **discrete coordination** $\mathbf{A}$ is a tuple $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_d)$ where each $\mathbf{A}_i$ is itself a tuple of real numbers in increasing order $\mathbf{A}_i = (\mathbf{A}_{i,1}, \ldots, \mathbf{A}_{i,n_i})$ such that $\mathbf{A}_{i,k+1} > \mathbf{A}_{i,k}$. These represent the coordinates of axis-separators along each of the d coordinate axes (Definition 13).

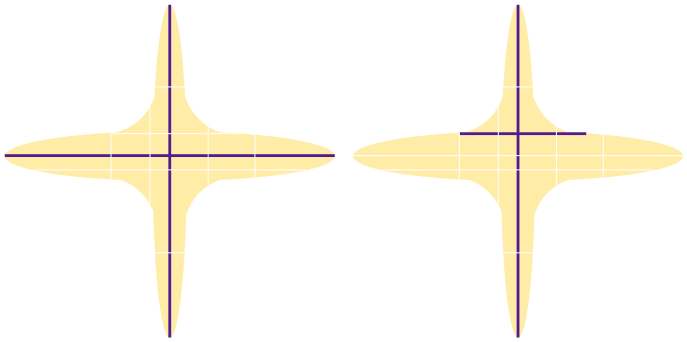


Figure 11: **Left**: The set of axis separators in dark blue are a backbone since they intersect all the others (Definition 14). **Right**: The set of separators in dark blue is **not** a backbone since the horizontal axis separator does not intersect the right-most vertical separator.

Appendix E. Detailed Proofs

E.1. Proof of non-removable discontinuity preservation (Theorem 4)

Proof Let us denote $J_h(z) = \frac{\partial h}{\partial z}(z)$ the Jacobian of h, and $J_{h^{-1}}(z') = \frac{\partial h^{-1}}{\partial z'}(z')$ the Jacobian of h^{-1}. Suppose p_Z is one of the PDFs of Z, from this we can obtain a PDF of Z' using the change of

variable formula: $p_{Z'}(z') = p_Z(h^{-1}(z'))|\det J_{h^{-1}}(z')|$. Symmetrically, we can say that if $p_{Z'}$ is a PDF of Z', we obtain a PDF version of Z as follows: $p_Z(z) = p_{Z'}(h(z))|\det J_h(z)|$.

Suppose the PDF p_Z has a non-removable discontinuity at z_0. Pick one of the PDFs of Z', let us call it $p_{Z'}$. There are three possibilities for what could happen at $Z' = h(z_0)$.

- $p_{Z'}$ is continuous at $h(z_0)$. We can apply the change of variables formula and obtain a PDF of Z that is given as $p_Z(z) = p_{Z'}(h(z))|\det J_h(z)|$. Since the RHS is a product of two terms that are continuous at z_0, we conclude that p_Z is continuous at z_0. This contradicts the fact that p_Z has a non-removable discontinuity at z_0.
- $p_{Z'}$ is discontinuous at $h(z_0)$ but the discontinuity is removable. Therefore, there exists a PDF $p_{Z'}$ that is continuous at $h(z_0)$. We can now follow the same argument as the above bullet to construct a PDF of Z that is continuous at z_0, which would contradict the fact that p_Z has a removable discontinuity at z_0.
- Finally, we are only left with the case that $p_{Z'}$ has a non-removable discontinuity at $h(z_0)$, which is what we set out to prove.

■

E.2. Theorem: a connected independent discontinuity of a PDF that has a connected support is an axis-separator of that support.

Theorem 18 *Let $\mathcal{S} \subset \mathbb{R}^d$ an open connected set. Let $\Gamma_{\mathcal{S}}(i,\tau) = \{z \in \mathcal{S} | z_i = \tau\}$. If $\Gamma_{\mathcal{S}}(i,\tau)$ is a connected set, then it is a separator of $\mathcal{S}$.*

E.3. Proof of grid structure preservation and recovery (Theorem 15)

E.3.1. STEP 1 – RECOVERY OF ALL SEPARATORS

Knowing, from Theorem 4, that the set of *points* making up the axis-aligned grid G maps through h to the set of *points* making up the axis-aligned grid G' (i.e. $G' = h(G)$), our first major step consists in establishing that the *axis-separators* that make up G (i.e. the elements of $\mathcal{G}$) map one-to-one to the *axis-separators* that make up G' (i.e. the elements of $\mathcal{G}'$). We can denote this simply as $G' = h(G) \implies \mathcal{G}' = h(\mathcal{G})$.

PROOF FOR STEP 1

The high level proof is as follows (a complete detailed proof is provided in Appendix E.5):

- Since $H \in \mathcal{G}$ is a connected smooth hypersurface in $\mathcal{S}$, and diffeomorphisms map connected sets to connected sets and smooth hypersurfaces to smooth hypersurfaces, we get that $h(H)$ is a connected smooth hypersurface in $\mathcal{S}'$.
- From Theorem 4, we also know that $h(H) \subset G'$.
- Next, we establish that the only smooth connected hypersurfaces in $\mathcal{S}'$ that are included in G' are necessarily subsets of a single axis-separator of $\mathcal{G}'$. This is fundamentally due to the fact that a connected smooth hypersurface cannot spread from one separator of the grid to another along their orthogonal intersection, as it would no longer be smooth (having a "kink"), so it has to stay within a single separator.
- We conclude that $h(H)$ is necessarily a subset of a single axis-separator $H' \in \mathcal{G}'$.

- We then show that not only is $h(H)$ a subset of a single axis-separator $H' \in \mathcal{G}'$, but that it has to be that entire separator. Because from the previous point, reverse diffeomorphism h^{-1} must map back the would-be remaining part of H' (i.e. $H' \setminus h(H) \neq \emptyset$) to a subset of the same separator as it maps back $h(H)$, i.e. to H. But this leads to a contradiction, since that remaining part did not come from H initially. (See proof of Lemma 26 in Appendix).
- We have thus shown that $H \in \mathcal{G} \implies h(H) \in \mathcal{G}'$. It suffices to apply this result in the other direction using h^{-1} to establish the converse. We thus have a bijection: the one-to-one mapping we needed to prove. Which we can write succinctly as $\mathcal{G}' = h(\mathcal{G})$.

E.3.2. STEP 2 – RECOVERY OF PARTITION INTO SETS OF PARALLEL SEPARATORS

We have established in step 1 that we recover the set of all separators $\mathcal{G}' = h(\mathcal{G})$. Our next step is to recover its partition into subsets of parallel separators (each subset associated to an axis): $\mathcal{G}'^{(j)} = h(\mathcal{G}^{(i)})$ (with permutation $j = \sigma(i)$).

PROOF FOR STEP 2

Consider d separators forming a backbone of $\mathcal{G}$, recall that a backbone is constituted of d distinct axis-separators that intersect in a single point, i.e. $\mathcal{H}_1^* \in \mathcal{G}^{(1)}, \dots, \mathcal{H}_d^* \in \mathcal{G}^{(d)}$, $\bigcap_{i=1}^d \mathcal{H}_i^* = \{z^*\}$.
We have that $\forall j \neq i,\ \mathcal{H}_i^* \neq \mathcal{H}_j^* \implies \forall j \neq i,\ h(\mathcal{H}_i^*) \neq h(\mathcal{H}_j^*)$.
We also have that $\bigcap_{i=1}^d \mathcal{H}_i^* = \{z^*\} \implies \bigcap_{i=1}^d h(\mathcal{H}_i^*) = \{h(z^*)\}$ (as h is a bijection).
Moreover, we know from step 1 that $\mathcal{H}_i^* \in \mathcal{G} \implies h(\mathcal{H}_i^*) \in \mathcal{G}'$. In short, the $h(\mathcal{H}_1^*), \dots, h(\mathcal{H}_d^*)$ are d distinct separators, each an element of $\mathcal{G}'$, that intersect in a single point $h(z^*)$. The only sets of d *distinct* separators in $\mathcal{G}'$ that pass through a same point are d separators defined along each of the d different axes of $\mathcal{Z}' = \mathbb{R}^d$. Thus there exists a permutation σ such that for such backbone separators, $\mathcal{H}_i^* \in \mathcal{G}^{(i)} \implies h(\mathcal{H}_i^*) \in \mathcal{G}'^{(\sigma(i))}$.

Now consider any other separator $H \in \mathcal{G}^{(i)}$. From the definition of the backbone, we know that $H \cap \mathcal{H}_j^* \neq \emptyset, \forall j \neq i$.
This implies that $h(H) \cap h(\mathcal{H}_j^*) \neq \emptyset, \forall j \neq i$. The fact that $h(H)$ intersects a separator $h(\mathcal{H}_j^*) \in \mathcal{G}'^{(\sigma(j))}$ implies that it does not belong to parallel-separator-set $\mathcal{G}'^{(\sigma(j))}$. Thus $\forall j \neq i,\ h(H) \notin \mathcal{G}'^{(\sigma(j))}$. So there is just one parallel separator set left which $h(H)$ can belong to: $h(H) \in \mathcal{G}'^{(\sigma(i))}$.
In short, we have proved that $H \in \mathcal{G}^{(i)} \implies h(H) \in \mathcal{G}'^{(\sigma(i))}$.

Since distinct separators map to distinct separators, and each has to belong to exactly one of the $\mathcal{G}'^{(k)}$, this mapping is a bijection and we can write $H \in \mathcal{G}^{(i)} \iff h(H) \in \mathcal{G}'^{(\sigma(i))}$, or in short $h(\mathcal{G}^{(i)}) = \mathcal{G}'^{(j)}$ with $j = \sigma(i)$.

E.3.3. STEP 3 – RECOVERY OF COORDINATE ORDERING

The last step consists in showing that the ordering of the separators in a parallel-separators-set is preserved (up to possible order reversal).

PROOF FOR STEP 3

The gist of the proof is as follows (a complete detailed proof is provided in Appendix E.4):

We first establish that h preserves separators and halves. This follows directly form the preservation of inclusion, connectedness and set operations under diffeomorphisms. Then, we use the fact that inclusion defines a strict order relationship between positive halves associated to a coordination,

and similarly between negative halves. As inclusion is preserved by a diffeomorphism, this order relationship is preserved. We can use this to show that the order implied by $\mathbf{A}_i$ is either conserved, as is, in $\mathbf{B}_j$ (negative halves of coordination $\mathbf{A}$ being mapped to negative halves of $\mathbf{B}$) or simply reversed (negative halves of $\mathbf{A}$ are being mapped to positive halves of $\mathbf{B}$). This directly yields the result of the main Theorem (15).

E.4. Detailed proof for Step 3

PRELIMINARY LEMMA

Lemma 19 *Preservation of separator and halves under diffeomorphism: If h is a diffeomorphism and $\mathcal{C}$ is a separator of $\mathcal{S}$ that splits it in two halves $\mathcal{C}^+$ and $\mathcal{C}^-$, then $h(\mathcal{C})$ is a separator of $h(\mathcal{S})$ that splits it in two halves $h(\mathcal{C}^+)$ and $h(\mathcal{C}^-)$*
Formally:

$$\begin{aligned} &\mathcal{C} \subset \mathcal{S},\ \mathcal{C},\ \text{connected},\ \text{split}(\mathcal{S},\ \mathcal{C}) = \{\mathcal{C}^+, \mathcal{C}^-\} \\ \iff &h(\mathcal{C}) \subset h(\mathcal{S}),\ h(\mathcal{C}),\ \text{connected},\ \text{split}(h(\mathcal{S}),\ h(\mathcal{C})) = \{h(\mathcal{C}^+), h(\mathcal{C}^-)\} \end{aligned}$$

Proof This follows from preservation of inclusion, connectedness, and set operations (union, intersection, difference) under a diffeomorphism.
Formally: $\mathcal{C} \subset \mathcal{S} \implies h(\mathcal{C}) \subset h(\mathcal{S})$.
$\mathcal{C}^+$ and $\mathcal{C}^-$ being the connected components of $\mathcal{S} - \mathcal{C}$ implies that $\mathcal{C}^+$ and $\mathcal{C}^-$ are each connected, and that $\mathcal{S} \setminus \mathcal{C} = \mathcal{C}^+ \cup \mathcal{C}^-$, where $\mathcal{C}^+ \cup \mathcal{C}^-$ is not connected.
Each of $\mathcal{S}$,$\mathcal{C}$, $\mathcal{C}^+$, $\mathcal{C}^-$ connected $\implies$ Each of $\mathcal{S}, h(\mathcal{C}), h(\mathcal{C}^+), h(\mathcal{C}^-)$ connected.
$\mathcal{S} \setminus \mathcal{C} = \mathcal{C}^+ \cup \mathcal{C}^- \implies h(\mathcal{S}) \setminus h(\mathcal{C}) = h(\mathcal{C}^+) \cup h(\mathcal{C}^-)$
$\mathcal{C}^+ \cup \mathcal{C}^-$ not connected $\implies$ $h(\mathcal{C}^+) \cup h(\mathcal{C}^-)$ not connected.
That $h(\mathcal{C}^+) \cup h(\mathcal{C}^-)$ is not connected but $h(\mathcal{C}^+)$ and $h(\mathcal{C}^-)$ are each connected, implies that $h(\mathcal{C}^+)$ and $h(\mathcal{C}^-)$ are the two connected components of $h(\mathcal{C}^+) \cup h(\mathcal{C}^-)$ i.e. of $h(\mathcal{S}) - h(\mathcal{C})$.
This implies that $\text{split}(h(\mathcal{S}), h(\mathcal{C})) = \{h(\mathcal{C}^+), h(\mathcal{C}^-)\}$. The implication in the other direction can be obtained in the by applying the same reasoning using h^{-1}. ■

PROOF OF STEP 3

Let $j = \sigma(i)$ and $K = |\mathbf{A}_i| = |\mathbf{B}_j|$ and denote the corresponding set of axis separators as
$\mathcal{A} = \{\Gamma_{\mathcal{S}}(i, \mathbf{A}_{i,1}), \ldots, \Gamma_{\mathcal{S}}(i, \mathbf{A}_{i,K})\}$ and $\mathcal{B} = \{\Gamma_{\mathcal{S}'}(j, \mathbf{B}_{j,1}), \ldots, \Gamma_{\mathcal{S}'}(j, \mathbf{B}_{j,K})\}$
and denote the corresponding sets of halves:
$\mathcal{A}^+ = \{\Gamma^+_{\mathcal{S}}(i, \mathbf{A}_{i,1}), \ldots, \Gamma^+_{\mathcal{S}}(i, \mathbf{A}_{i,K})\}$, $\mathcal{A}^- = \{\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,1}), \ldots, \Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,K})\}$, $\mathcal{A}^\pm = \mathcal{A}^+ \cup \mathcal{A}^-$
and $\mathcal{B}^+ = \{\Gamma^+_{\mathcal{S}'}(j, \mathbf{B}_{j,1}), \ldots, \Gamma^+_{\mathcal{S}'}(j, \mathbf{B}_{j,K})\}$, $\mathcal{B}^- = \{\Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,1}), \ldots, \Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,K})\}$, $\mathcal{B}^\pm = \mathcal{B}^+ \cup \mathcal{B}^-$
Proof Step 2, states that $h(\mathcal{A}) = \mathcal{B}$.
And we have from the above Lemma that

$$\begin{aligned} &\text{split}(\mathcal{S}, \mathcal{C}) = \{\mathcal{C}^+, \mathcal{C}^-\} \\ \iff &\text{split}(h(\mathcal{S}), h(\mathcal{C})) = \{h(\mathcal{C}^+), h(\mathcal{C}^-)\} \end{aligned}$$

thus the equality of the sets of separators $h(\mathcal{A}) = \mathcal{B}$ obtained in Proof Step 2 implies an equality of the sets of halves:

$$h(\mathcal{A}^{\pm}) = \mathcal{B}^{\pm}$$

Now, the only halves, among all halves, that do not include any of the separators are $\Gamma^{-}_{\mathcal{S}}(i, \mathbf{A}_{i,1})$ and $\Gamma^{+}_{\mathcal{S}}(i, \mathbf{A}_{i,K})$ i.e. formally:

$$\{\mathcal{C} \in \mathcal{A} | \forall \mathcal{H} \in \mathcal{A}^{\pm}, \mathcal{C} \cap \mathcal{H} = \emptyset\} = \{\Gamma^{-}_{\mathcal{S}}(i, \mathbf{A}_{i,1}), \Gamma^{+}_{\mathcal{S}}(i, \mathbf{A}_{i,K})\}$$

this property will naturally translate to their mapping by diffeomorphism h (due to preservation of inclusion an intersections)

hence

$$\{\mathcal{C} \in h(\mathcal{A}) | \forall \mathcal{H} \in h(\mathcal{A}^{\pm}), \mathcal{C} \cap \mathcal{H} = \emptyset\} = \{h(\Gamma^{-}_{\mathcal{S}}(i, \mathbf{A}_{i,1})), h(\Gamma^{+}_{\mathcal{S}}(i, \mathbf{A}_{i,K}))\}$$

i.e.

$$\{\mathcal{C} \in \mathcal{B} | \forall \mathcal{H} \in \mathcal{B}^{\pm}, \mathcal{C} \cap \mathcal{H} = \emptyset\} = \{h(\Gamma^{-}_{\mathcal{S}}(i, \mathbf{A}_{i,1})), h(\Gamma^{+}_{\mathcal{S}}(i, \mathbf{A}_{i,K}))\}$$

but we also have, similarly,

$$\{\mathcal{C} \in \mathcal{B} | \forall \mathcal{H} \in \mathcal{B}^{\pm}, \mathcal{C} \cap \mathcal{H} = \emptyset\} = \{\Gamma^{-}_{\mathcal{S}'}(i, \mathbf{B}_{i,1})), \Gamma^{+}_{\mathcal{S}'}(i, \mathbf{B}_{i,K}))\}$$

From this we conclude that:

$$\{h(\Gamma^{-}_{\mathcal{S}}(i, \mathbf{A}_{i,1})), h(\Gamma^{+}_{\mathcal{S}}(i, \mathbf{A}_{i,K}))\} = \{\Gamma^{-}_{\mathcal{S}'}(i, \mathbf{B}_{i,1})), \Gamma^{+}_{\mathcal{S}'}(i, \mathbf{B}_{i,K}))\}$$

Thus we have either one of two cases:

Case 1: $h(\Gamma^{-}_{\mathcal{S}}(i, \mathbf{A}_{i,1})) = \Gamma^{-}_{\mathcal{S}'}(i, \mathbf{B}_{i,1})$ and $h(\Gamma^{+}_{\mathcal{S}}(i, \mathbf{A}_{i,K})) = \Gamma^{+}_{\mathcal{S}'}(i, \mathbf{B}_{i,K})$. We associate this case with $s_i = +1$

Case 2: $h(\Gamma^{-}_{\mathcal{S}}(i, \mathbf{A}_{i,1})) = \Gamma^{+}_{\mathcal{S}'}(i, \mathbf{B}_{i,K})$ and $h(\Gamma^{+}_{\mathcal{S}}(i, \mathbf{A}_{i,K})) = \Gamma^{-}_{\mathcal{S}'}(i, \mathbf{B}_{i,1})$. We associate this case with $s_i = -1$

Case 1: $s_i = +1$**,** $h(\Gamma^{-}_{\mathcal{S}}(i, \mathbf{A}_{i,1})) = \Gamma^{-}_{\mathcal{S}'}(i, \mathbf{B}_{i,1})$ **and** $h(\Gamma^{+}_{\mathcal{S}}(i, \mathbf{A}_{i,K})) = \Gamma^{+}_{\mathcal{S}'}(i, \mathbf{B}_{i,K})$ The half-spaces in $\mathcal{A}^{\pm}$ that include $\Gamma^{-}_{\mathcal{S}}(i, \mathbf{A}_{i,1})$ are only the $\Gamma^{-}_{\mathcal{S}}$, formally:

$$\{\mathcal{H} \in \mathcal{A}^{\pm} | \Gamma^{-}_{\mathcal{S}}(i, \mathbf{A}_{i,1}) \subset \mathcal{H}\} = \mathcal{A}^{-}$$

this relationship will be maintained under diffeomorphism h i.e.

$$\{\mathcal{H} \in h(\mathcal{A}^{\pm}) | h(\Gamma^{-}_{\mathcal{S}}(i, \mathbf{A}_{i,1})) \subset \mathcal{H}\} = h(\mathcal{A}^{-})$$

thus, since $h(\mathcal{A}^{\pm}) = \mathcal{B}^{\pm}$ and $h(\Gamma^{-}_{\mathcal{S}}(i, \mathbf{A}_{i,1})) = \Gamma^{-}_{\mathcal{S}'}(i, B_{i,1})$ this can be rewritten as

$$\{\mathcal{H} \in \mathcal{B}^{\pm} | \Gamma^{-}_{\mathcal{S}'}(i, \mathbf{B}_{i,1}) \subset \mathcal{H}\} = h(\mathcal{A}^{-})$$
$$\mathcal{B}^{-} = h(\mathcal{A}^{-})$$

or, written less compactly:

$$\{h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,1})), \ldots, h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,K}))\} = \{\Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,1}), \ldots, \Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,K})\}$$

Furthermore strict inclusion defines an order relationship between the elements of $\mathcal{A}^-$ which will be preserved under the diffeomorphism, and thus defines a strict ordering between them:

$$\begin{aligned} &\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,1}) \subsetneq \Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,2}) \subsetneq \ldots \subsetneq \Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,K}) \\ \implies &h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,1})) \subsetneq h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,2})) \subsetneq \ldots \subsetneq h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,K})) \end{aligned}$$

we know that the $h(\Gamma^-_{\mathcal{S}}(i, A_{i,k}))$ are the elements of $\mathcal{B}^-$ (as we have just sown that $\mathcal{B}^- = h(\mathcal{A}^-)$), i.e. the $\Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,k})$. Their order is defined uniquely by strict inclusion as

$$\Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,1}) \subsetneq \Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,2}) \subsetneq \ldots \subsetneq \Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,K})$$

thus we can conclude not only (as we showed with $\mathcal{B}^- = h(\mathcal{A}^-)$) that

$$\{h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,1})), \ldots, h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,K}))\} = \{\Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,1}), \ldots, \Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,K})\}$$

but also that their ordering is preserved i.e.

$$(h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,1})), \ldots, h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,K}))) = (\Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,1}), \ldots, \Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,K}))$$

or expressed differently:

$$\forall k \in \{1, \ldots, K\}, h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,k})) = \Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,k})$$

it is straightforward to conclude from this that we also have

$$\begin{aligned} &\forall k \in \{1, \ldots, K\}, \\ &h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,k})) = \Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,k}) \\ &h(\Gamma^+_{\mathcal{S}}(i, \mathbf{A}_{i,k})) = \Gamma^+_{\mathcal{S}'}(j, \mathbf{B}_{j,k}) \\ &h(\Gamma_{\mathcal{S}}(i, \mathbf{A}_{i,k})) = \Gamma_{\mathcal{S}'}(j, \mathbf{B}_{j,k}) \end{aligned}$$

or stated differently, that:

$$\begin{aligned} &\forall k \in \{1, \ldots, K\}, \forall z \in \mathcal{S} \\ &z \in \Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,k}) \iff h(z) \in \Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,k}) \\ &z \in \Gamma^+_{\mathcal{S}}(i, \mathbf{A}_{i,k}) \iff h(z) \in \Gamma^+_{\mathcal{S}'}(j, \mathbf{B}_{j,k}) \\ &z \in \Gamma_{\mathcal{S}}(i, \mathbf{A}_{i,k}) \iff h(z) \in \Gamma_{\mathcal{S}'}(j, \mathbf{B}_{j,k}) \end{aligned}$$

or equivalently

$$\begin{aligned}
&\forall k \in \{1, \dots, K\}, \forall z' \in \mathcal{S}', \\
&h^{-1}(z') \in \Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,k}) \iff z' \in \Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,k}) \\
&h^{-1}(z') \in \Gamma^+_{\mathcal{S}}(i, \mathbf{A}_{i,k}) \iff z' \in \Gamma^+_{\mathcal{S}'}(j, \mathbf{B}_{j,k}) \\
&h^{-1}(z') \in \Gamma_{\mathcal{S}}(i, \mathbf{A}_{i,k}) \iff z' \in \Gamma_{\mathcal{S}'}(j, \mathbf{B}_{j,k})
\end{aligned}$$

which we may also write

$$\begin{aligned}
&\forall k \in \{1, \dots, K\}, \forall z' \in \mathcal{S}', \\
&z'_j < \mathbf{B}_{j,k} \iff h^{-1}(z')_i < \mathbf{A}_{i,k} \\
&z'_j > \mathbf{B}_{j,k} \iff h^{-1}(z')_i > \mathbf{A}_{i,k} \\
&z'_j = \mathbf{B}_{j,k} \iff h^{-1}(z')_i = \mathbf{A}_{i,k}
\end{aligned}$$

which is what we needed to prove in the main grid structure recovery theorem.

Case 2: axis reversal $s_i = -1$**,** $h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,1})) = \Gamma^+_{\mathcal{S}'}(i, \mathbf{B}_{i,K})$ **and** $h(\Gamma^+_{\mathcal{S}}(i, \mathbf{A}_{i,K})) = \Gamma^-_{\mathcal{S}'}(i, \mathbf{B}_{i,1})$
We can follow the exact same reasoning steps as in case 1, starting from $h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,1})) = \Gamma^+_{\mathcal{S}'}(i, \mathbf{B}_{i,K})$:

- to first show that $h(\mathcal{A}^-) = \mathcal{B}^+$ i.e.

$$\{h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,1})), \dots, h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,K}))\} = \{\Gamma^+_{\mathcal{S}'}(j, \mathbf{B}_{j,1}), \dots, \Gamma^+_{\mathcal{S}'}(j, \mathbf{B}_{j,K})\}$$

- then use the preservation of the order relation defined by inclusion of halves to establish that

$$(h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,1})), \dots, h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,K}))) = (\Gamma^+_{\mathcal{S}'}(j, \mathbf{B}_{j,K}), \dots, \Gamma^+_{\mathcal{S}'}(j, \mathbf{B}_{j,1}))$$

- thus that

$$\begin{aligned}
&\forall k \in \{1, \dots, K\}, \\
&h(\Gamma^-_{\mathcal{S}}(i, \mathbf{A}_{i,k})) = \Gamma^+_{\mathcal{S}'}(j, \mathbf{B}_{j,K-k+1}) \\
&h(\Gamma^+_{\mathcal{S}}(i, \mathbf{A}_{i,k})) = \Gamma^-_{\mathcal{S}'}(j, \mathbf{B}_{j,K-k+1}) \\
&h(\Gamma_{\mathcal{S}}(i, \mathbf{A}_{i,k})) = \Gamma_{\mathcal{S}'}(j, \mathbf{B}_{j,K-k+1})
\end{aligned}$$

- conclude that

$$\begin{aligned}
&\forall k \in \{1, \dots, K\}, \forall z' \in \mathcal{S}', \\
&\quad z'_j > \mathbf{B}_{j,k} \iff h^{-1}(z')_i < \mathbf{A}_{i,K-k+1} \\
&\quad z'_j < \mathbf{B}_{j,k} \iff h^{-1}(z')_i > \mathbf{A}_{i,K-k+1} \\
&\quad z'_j = \mathbf{B}_{j,k} \iff h^{-1}(z')_i = \mathbf{A}_{i,K-k+1}
\end{aligned}$$

which is what we needed to prove in the main grid structure recovery theorem.

E.5. Detailed proof of Step 1 – recovery of all separators

The goal of step 1 is to establish that the *axis-separators* that make up G (i.e. the elements of $\mathcal{G}$) map one-to-one to the *axis-separators* that make up G' (i.e. the elements of $\mathcal{G}'$). We can denote this simply as $G' = h(G) \Longrightarrow \mathcal{G}' = h(\mathcal{G})$.

A succinct overview of of the proof was given in Section E.3.1 in the main text. We provide a detailed proof here. Note that we always assume finite axis-separator sets.

PRELIMINARIES

Whenever we say hypersurface, it is always defined as a $d-1$ dimensional regular submanifold embedded in d-dimensional *ambient space* $\mathcal{S} \subset \mathbb{R}^d$, where $\mathcal{S}$ is a d-dimensional connected open submanifold of $\mathbb{R}^d$. In our application $\mathcal{S}$ will be the interior of the support of the density we consider.

- **Definition: Intersection set.** given a grid $G = \cup\mathcal{G} = \cup_{H\in\mathcal{G}} H$, we define its *intersection set* $I(\mathcal{G})$ as the set of points that belong to intersections of 2 or more distinct separators of $\mathcal{G}$. Formally: $I(\mathcal{G}) = \cup_{H\in\mathcal{G},H'\in\mathcal{G},H'\neq H}(H \cap H')$.
- **Definition: Exclusive point.** We say that a point z is exclusive to a separator H of a grid $G = \cup\mathcal{G}$ if it belongs to H but does not belong to any other separator of the grid (i.e. it does not belong to I). Similarly we will say that a set is exclusive to a separator if all its elements are exclusive points of that separator. The set of points of a separator H that are exclusive to it will be denoted $\breve{H} = H \setminus I$.
- **Definition: Tangent space** we view the tangent spaces to hypersurfaces embedded in an ambient space included in $\mathbb{R}^d$ literally as affine subspaces of $\mathbb{R}^d$, i.e. we use the traditional view[5] of tangent space (do Carmo, 1976), which is a natural generalization of the notion of a plane tangent to a surface at a point, to higher dimensional hypersurfaces embedded in $\mathbb{R}^d$. The tangent space at $z \in A$ to a hypersurface A will be denoted $T^A(z) = T_z A$. A smooth hypersurface has the property that it has at every $z \in A$ a well-defined tangent space $T^A(z) = T_z A$ of the same dimension as the hypersurface. When A is a smooth hypersurface, T^A is a smooth map $T^A : A \to \mathrm{Graff}_{d-1}(\mathbb{R}^d)$ that maps any point z of A to a point of the *affine-Grassmannian manifold* (Klain and Rota, 1997; Lim et al., 2021) $\mathrm{Graff}_{d-1}(\mathbb{R}^d)$, i.e. the space of all $d-1$ dimensional affine-subspaces of $\mathbb{R}^d$. Since T^A is a continuous map between smooth manifolds, $T^A(z)$ will be continuous in any local (or global) parametrization of A around z. (continuity based on the topology of the affine-Grassmannian manifold for comparing tangent spaces as affine-subspaces of $\mathbb{R}^d$).
  Note that the tangent space to any axis-separator H is a constant: it is the subspace confounded with the hyperplane that includes the separator, and will be denoted $\mathcal{T}_H$. i.e. we have $\forall z \in H,\ T^H(z) = T_z H = \mathcal{T}_H$. Note also that with this affine subspace definition of tangent space, $\mathcal{T}_H$ is different for every separator H of an axis-aligned grid: $\forall H_1 \in \mathcal{G}, \forall H_2 \in \mathcal{G}, \mathcal{T}_{H_1} = \mathcal{T}_{H_2} \Leftrightarrow H_1 = H_2$.
- **Useful properties:** We will also use the following properties that are either well-established differential geometry knowledge or straightforward corollaries thereof
  - **Property 1:** *A diffeomorphism maps a smooth hypersurface to a smooth hypersurface*
  - **Property 2:** *A diffeomorphism maps a path-connected set to a path-connected set.*

5. This traditional extrinsic view of tangent space is preferred here to more modern definitions, because it simplifies a step in our proof. It is also arguably easier to intuit and follow for readers who may not be familiar with differential geometry.

- **Property 3:** *Smooth connected hypersurfaces in $\mathbb{R}^d$ have a $d-1$ dimensional tangent space that is well-defined all over the hypersurface and continuous (in the sense defined above, see Tangent space)*
- **Property 4:** *A non-empty open subset of a smooth hypersurface in ambient space is itself a smooth hypersurface in ambient space*
- **Property 5:** *A hypersurface that is a subset of another hypersurface has at every of its points the same tangent space as the hypersurface it is a subset of.*

DETAILED PROOF OF STEP 1

Lemma 20 *No subset of the intersection set $I(\mathcal{G})$ of an axis-aligned grid $G = \cup\mathcal{G}$ can be a hypersurface in ambient space.*

Proof Consider $I(\mathcal{G})$ the intersection set of a grid $G = \cup\mathcal{G}$. Formally:
$I(\mathcal{G}) = \cup_{H\in\mathcal{G},H'\in\mathcal{G},H'\neq H}(H \cap H')$. Each $H \cap H'$, if it is non-empty, is the intersection of two orthogonal (thus transversal) connected hypersurfaces (i.e. $d-1$ dimensional submanifolds embedded in ambient space), so that their intersection can be at most a $d-2$ dimensional embedded submanifold of ambient space. The union of a finite number of at most $d-2$ dimensional submanifolds cannot be more than $d-2$ dimensional, so $I(\mathcal{G})$ cannot be more than $d-2$ dimensional. Consequently no subset of $I(\mathcal{G})$ can be more than $d-2$ dimensional, thus it cannot be a hypersurface in ambient space. ■

Lemma 21 *Let A be a connected smooth hypersurface included in an axis-alined grid $G = \cup\mathcal{G}$ with axis-separator set $\mathcal{G}$. Let $z \in A$. All open neighborhoods of z in A will necessarily contain at least one point that is exclusive to a separator of $\mathcal{G}$.*

Proof An open neighborhood $\mathcal{B}_z^A$ of z in A is an open subset of A, thus from Property 4, $\mathcal{B}_z^A$ is a hypersurface in ambient space. From Lemma 20 no subset of $I(\mathcal{G})$, (the set of points of G that belong to more than one separator) can be a hypersurface. So $\mathcal{B}_z^A$ cannot be a subset of $I(\mathcal{G})$, i.e. it must contain at least one point exclusive to a separator of $\mathcal{G}$. ■

Lemma 22 *Let A be a connected smooth hypersurface included in an axis-alined grid $G = \cup\mathcal{G}$ with axis-separator set $\mathcal{G}$. Let z be a point of A that is exclusive to a separator $H \in \mathcal{G}$ (i.e. $z \in H \setminus I(\mathcal{G})$: it belongs to no other separator of $\mathcal{G}$), then there exists an open connected neighborhood $\mathcal{B}_z^A$ of z in A that is exclusive to H.*

Proof We reason using the usual Euclidean distance in $\mathbb{R}^d$. Consider an open d-ball $\mathcal{B}_z^d$ in $\mathbb{R}^d$ centered on z and whose radius ϵ is chosen to be less than the smallest distance of z to any other separator, i.e. such that $0 < \epsilon < \inf_{z'\in(G\setminus H)} \|z - z'\|$. Since z is exclusive to separator H and the number of separators is finite, this distance will be greater than 0. Then all points of G within a distance less than ϵ of z will necessarily belong exclusively to H, i.e. $\mathcal{B}_z^d \cap G \subset \breve{H}$, where $\breve{H} = H \setminus I(\mathcal{G})$. Now we can choose a sufficiently small connected open neighborhood $\mathcal{B}_z^A$ of z in A so that the distance in ambient space between z and any other point of $\mathcal{B}_z^A$ is less than ϵ. Thus $\mathcal{B}_z^A \subset \mathcal{B}_z^d$. Since we also have $\mathcal{B}_z^A \subset A \subset G$ this implies that $\mathcal{B}_z^A \subset \mathcal{B}_z^d \cap G$ and consequently that $\mathcal{B}_z^A \subset \breve{H}$. We have thus shown that there exists an open connected neighborhood of z in A that is exclusive to H. ■

Lemma 23 *Let A be a connected smooth hypersurface included in an axis-alined grid $G = \cup\mathcal{G}$ with axis-separator set $\mathcal{G}$. Then for any point $z \in A$ there exists a non-empty open subset B whose boundary contains z and such that B is a non-empty open subset exclusive to one of the separators.*

Proof There are two cases to consider for z: either z is an exclusive point of a separator of the grid, or it is an intersection point of separators (belonging to $I(\mathcal{G})$).

First case: z is a point exclusive to a separator $H \subset \mathcal{G}$.
Then, by Lemma 22, we know that there exists an open connected neighborhood $\mathcal{B}_z^A$ of z in A that is exclusive to H. We can then easily pick an open subset B of $\mathcal{B}_z^A$ whose boundary contains z (For instance, pick a close neighbor z_1 of z in $\mathcal{B}_z^A$, and construct B as the intersection of $\mathcal{B}_z^A$ with an open ball centered on z_1 and of radius $\|z_1 - z\|$). B is an open subset exclusive to H, the separator that z belongs to.

Second case: z is not exclusive to any separator of the grid.
Let $\mathcal{G} = \{H_1, \dots, H_k\}$ be the finite set of separators of grid $G = \cup\mathcal{G}$. Let $\breve{H}_i = H_i \setminus I(\mathcal{G})$ be the corresponding subset of exclusive points to each separator H_i, and let $\breve{A}_i = \breve{H}_i \cap A$, for each $i \in \{1, \dots, k\}$. So $\breve{A}_i$, if it is not empty, will contain only points exclusive to H_i. From Lemma 22 we deduce that every point of $\breve{A}_i$ has an open neighborhood in A exclusive to H_i: this open neighborhood is thus included in $A \cap \breve{H}_i$ and is thus a subset of $\breve{A}_i$. We have thus shown that every point of $\breve{A}_i$ has an open neighborhood in A that is included in $\breve{A}_i$. From this we conclude that each $\breve{A}_i$ is an open subset (possibly empty) of A.
We know that z belongs to none of the $\breve{A}_i$, since it is not exclusive to any separator. Now we will show that z belongs to the *boundary* of at least one of the $\breve{A}_i$. We will reason using the metric d^A induced on embedded submanifold $A \subset \mathbb{R}^d$ by the usual Euclidean metric in ambient space $\mathbb{R}^d$. Let $\epsilon = \min_{i\in\{1,\dots,k\}} d^A(z, \breve{A}_i)$. We can use a min since it is over a finite number k of separators. Note that $d^A(z, \breve{A}_i) = \inf_{z'\in\breve{A}_i} d^A(z, z')$ will be $+\infty$ if $\breve{A}_i$ is empty, by the definition of the infimum. If ϵ was strictly greater than 0, then this would mean that no point of A exclusive to any separator would be at a distance strictly less than ϵ from z (since any point of A exclusive to a separator belongs to one of the $\breve{A}_i$). Thus the open ball $\mathcal{B}^A(z, \epsilon) = \{z' \in A, d^A(z, z') < \epsilon\}$ would not contain any point exclusive to any separator. But this would contradict Lemma 21. So necessarily $\epsilon = 0$. This implies that there is at least one of the $\breve{A}_i$ whose distance to z is 0, i.e. there exists a $k^* \in \{0, \dots, k\}$ such that $d^A(z, \breve{A}_{k^*}) = 0$. Since $z \notin \breve{A}_{k^*}$ we conclude that z belongs to the *boundary* of this $\breve{A}_{k^*}$. Moreover this $\breve{A}_{k^*}$ is non-empty (otherwise that distance would be $+\infty$). It is thus an open-subset of A, exclusive to separator H_k. We have thus established that there exists a non-empty open subset of A exclusive to one of the separators, and whose boundary contains z.

■

Lemma 24 *Let A be a connected smooth hypersurface included in an axis-alined grid $G = \cup\mathcal{G}$ with axis-separator set $\mathcal{G}$. Let γ be a continuous path, included in A, that starts at a point z_1, where z_1 is exclusive to a separator $H \in \mathcal{G}$. Then γ will necessarily be included entirely in H.*

Proof Consider path $\gamma : [0, 1] \to A$, where $\gamma(0) = z_1$ is exclusive to H. From Lemma 23, for each point $\gamma(t) \in A$, there exists an open subset B_t of A whose boundary contains $\gamma(t)$ and such that B_t is an open subset exclusive to one of the separators. Let us call this separator H_t (Note that there may be multiple possible choices for B_t and H_t). Consider any point $z \in B_t$. Since B_t is a non-empty open subset of smooth hypersurface A, by Property 4 it is a hypersurface and by Property 5 B_t

and A will have the same tangent space, so that $T_z B_t = T_z A$. Since B_t is also a subset of smooth hypersurface H_t, we have by Property 5 that $T_z B_t = T_z H_t$. Thus $T_z A = T_z B_t = T_z H_t$. Now the tangent space to any axis-separator H' is the constant $\mathcal{T}_{H'}$. We can thus write, for any $z \in B_t$, $T_z A = T_z B_t = T_z H_t = \mathcal{T}_{H_t}$. Since A is a connected smooth hypersurface, it has a continuous and well defined tangent space at every point. Thus if the tangent space is constant on an open subset $B_t \subset A$ it will have that same constant value at its boundary. So the tangent space to A at point $\gamma(t)$, which belongs to the boundary of B_t, will also be $T_{\gamma(t)} A = \mathcal{T}_{H_t}$. For the same reason of the continuity of the tangent space of a path-connected smooth hypersurface A, we cannot have, along the curve $\gamma(t)$, an abrupt change in the tangent space $T_{\gamma(t)} A$, consequently $\mathcal{T}_{H_t}$ cannot change abruptly along the path. The only way for it not to change abruptly is that H_t stays constant along the path: $H_t = \text{constant } \forall t \in [0, 1]$. In other words, for any point $\gamma(t)$ along the path there must exist an open subset B_t of A that is included in and exclusive to the same constant separator along the path. Now, if $z_1 = \gamma(0)$ is exclusive to a separator H, then $H_0 = H$ and we must thus have $H_t = H$, $\forall t \in [0, 1]$. We have thus shown that if the path starts at a point $z_1 = \gamma(0)$ which is exclusive to a separator H, then all points of the path necessarily belong to H (though not necessarily exclusively to H). Thus, the path is entirely included in H. ■

Lemma 25 *A path-connected smooth hypersurface A included in an axis-alined grid $G = \cup\mathcal{G}$ is necessarily a subset of one separator of $\mathcal{G}$.*

Proof Let z_0 be a point of A that is exclusive to a separator $H \in \mathcal{G}$. We know from Lemma 21 that such a point exists. Since A is path-connected, there exists in A a continuous path connecting z_0 to any point $z \in A$. Thus, Lemma 24 leads to conclude that $\forall z \in A, z \in H$. Thus $A \subset H$. ■

Lemma 26 Let $G = \cup\mathcal{G}$ be an axis-aligned grid in $\mathcal{S}$. Let $h : \mathcal{S} \to \mathcal{S}'$ be a diffeomorphism. Let $G' = \cup\mathcal{G}'$ be an axis-aligned grid in $\mathcal{S}'$. If $h(G) = G'$, the image of a separator $H_1 \in \mathcal{G}$ by the diffeomorphism h will be a separator $H' \in \mathcal{G}'$, i.e. $H_1 \in \mathcal{G} \implies h(H_1) \in \mathcal{G}'$.

Proof An axis-separator $H_1 \subset G$ is a path-connected smooth hypersurface. From Property 1 and Property 2, its image by diffeomorphism h will be a path-connected smooth hypersurface $h(H_1) \subset G'$. So $h(H_1)$ is a path-connected smooth hypersurface included in axis-aligned grid G'. Consequently, Lemma 25 guarantees that we have $h(H_1) \subset H'$ for some $H' \in \mathcal{G}'$. We will now prove that $h(H_1) = H'$. Suppose by contradiction that $h(H_1) \subsetneq H'$, and let $B' = H' - h(H_1) \neq \emptyset$. Similarly, if we apply the reverse diffeomorphism h^{-1}, we will have $h^{-1}(H') \subset H_2$ for some $H_2 \in \mathcal{G}$. Consequently, the two disjoint sets composing $H' = B' \cup h(H_1)$ will both map back to subsets of H_2, i.e. $h^{-1}(B') \subset H_2$ and $h^{-1}(h(H_1)) \subset H_2$. The latter can be rewritten as $H_1 \subset H_2$, which implies $H_2 = H_1$ since no two distinct separators of $\mathcal{G}$ are included in one another. So we have $h^{-1}(B') \subset H_1$. Thus $h(h^{-1}(B')) \subset h(H_1)$, hence $B' \subset h(H_1)$. We had defined B' as $B' = H' - h(H_1) \neq \emptyset$ but a non-empty B' cannot at the same time correspond to a set from which we removed $h(H_1)$ *and* be included in $h(H_1)$. We have a contradiction, so we cannot have $h(H_1) \subsetneq H'$, therefore $h(H_1) = H'$. ■

Proposition 27 *The diffeomorphism h maps separators in $\mathcal{G}$ one-to-one to separators in $\mathcal{G}'$, i.e. $h(G) = G' \implies h(\mathcal{G}) = \mathcal{G}'$.*

Proof We have shown in Lemma 26 that $H \in \mathcal{G} \implies h(H) \in \mathcal{G}'$. It suffices to apply this result in the other direction using h^{-1} to establish the converse. We thus have a bijection: the one-to-one mapping we needed to prove. Which we can write succinctly $h(\mathcal{G}) = \mathcal{G}'$. ■

Appendix F. Background on non-removable discontinuities

The definition of continuity of a function is leveraged in section 5.1.

Definition 28 *A function f is continuous at a point x_0 if*

$$\forall \epsilon > 0;\ \exists \delta > 0\ :\ d(x, x_0) < \delta\ \Rightarrow\ |f(x) - f(x_0)| < \epsilon$$

for x in the domain of f and $d(x, x_0)$ being the distance between points x and x_0.

That is, for any positive real number ϵ, which can be infinitely small, there exists a positive real number δ such that for x in the interval $x_0 - \delta < x < x_0 + \delta$, the function of x will be at the interval $f(x_0) - \epsilon < f(x) < f(x_0) + \epsilon$. So for $f(x)$ to be in a small neighborhood around $f(x_0)$, x can be chosen in a small neighborhood around x_0.

Definition 29 *(Marsden and Hoffman, 1993) A function $f : A \subset M \to N$ is called **continuous on** the set $B \subset A$ if f is continuous at each point of B. If we just say that f is **continuous**, we mean that f is continuous on its domain A.*

These definitions are relevant because when a function is continuous, Definition 28 will hold at all the points in its domain. However, there can be cases where a function is not continuous at a point, but it is continuous almost everywhere. We define the **removable discontinuity** of a PDF p at point x_0 as a discontinuity that can be removed by mapping it to another PDF p' that has the same probability measure. The idea is that the area under the curve is the same in the equivalent PDF without the discontinuity at any interval, as illustrated in Figure 12. More precisely, the PDF p_x represents a probability distribution where we can evaluate the probability at an interval by integrating over it, such as $\Pr[a \leq X \leq b] = \int_a^b p_X(x)dx$ for a, b belonging to a *measurable set* $\mathcal{M}$ [6]. Hence, a probability distribution Q maps a measurable set $\mathcal{M}$ to $[0, 1]$. $Q : \mathcal{M} \to [0, 1]$. We notice and exemplify in the figure that multiple PDFs can represent the same probability distribution. When PDFs represent the same probability distribution, we will use the terminology that they belong to the same *equivalence class*.

6. We refer to (Capinski and Kopp, 2013) Chapter 2, definition 2.3, for coverage on measurable sets.

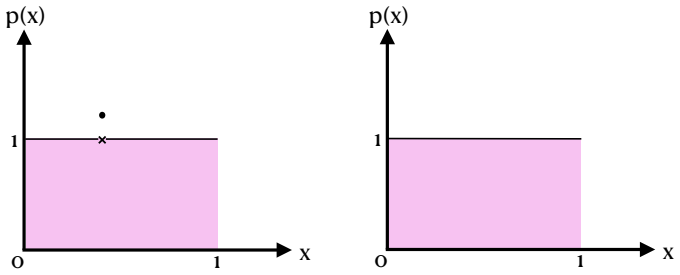


Figure 12: The PDF on the left has a removable discontinuity, but it can be mapped to the PDF on the right, which is identical but continuous everywhere. In whatever interval taken within the domain, the area under the PDF is exactly the same for both of them.

A **non-removable discontinuity** is the type of discontinuity that cannot be removed because the discontinuity affects the area below the PDF and therefore all the PDFs in the same equivalence class present a discontinuity at x_0, as illustrated in Figure 13.

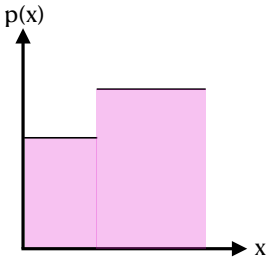


Figure 13: Example of a PDF with a non-removable discontinuity. The area under the PDF is affected by the discontinuity.

Appendix G. On the necessity of axis-aligned landmarks in the probability density p_Z

Our justification is built on a powerful result, Theorem 6 from Buchholz et al. (2022). We restate and revisit the intuitions behind the result below. We use the notation $\Phi_* \mathbb{P}$ to denote the pushforward of $\mathbb{P}$ under Φ.

Theorem 30 *Let p_Z be a twice differentiable probability density with bounded gradient. Suppose that $x = \Phi(z)$ where the distribution $\mathbb{P}$ of z has density p_Z and Φ is a diffeomorphism with $detD\Phi(z) = 1$ for $z \in \mathbb{R}^d$. Then there is a family of functions $\Phi_t : \times\mathbb{R}^d \to \mathbb{R}^d$ indexed by t with $\Phi_0 = q$ and $q_t \neq \Phi_0$ for $t \neq 0$ such that $detD\Phi_t(z) = 1$ and $(\Phi_t)_* \mathbb{P} = \Phi_* \mathbb{P}$.*

Consider the case when Φ is an identity map. In that case, the above theorem implies that there exists a family of volume preserving transformations different from the identity map such that $(\Phi_t)_* \mathbb{P} = \mathbb{P}$.

We revisit the intuition behind the proof, which will be reused for the rest of this section. We define the flow of a vector field as a map $\Phi : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\Phi_0(z) = z, \partial_t \Phi_t(z) = X(\Phi_t(z)) \tag{4}$$

We write $\Phi(t, z)$ as $\Phi_t(z)$ in the above expression. We can interpret $\Phi_t(z)$ as the position of a particle, which started at z at time $t = 0$. Then $X(\Phi_t(z))$ is the velocity of the particle at time t. Define a vector field $X^{ij} : \mathbb{R}^d \to \mathbb{R}^d$ as

$$X^{ij}_k = \begin{cases} \partial p_j & k = i \\ -\partial p_i & k = j \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

where X^{ij}_k is the k^{th} component of the vector field. Observe that X^{ij} is orthogonal to the isolines of the density p_Z along the plane corresponding to components i and j. Under the flows described in equation (4), the probability measures evolve in time and satisfy the continuity equation stated next. Formally stated, the density associated with $(q_t)_*\mathbb{P}$ satisfies.

$$\partial p_t + \mathsf{Div}(p_t X^{ij}) = 0, \quad p_0 = p \tag{6}$$

Observe that $\mathsf{Div}(X^{ij}) = \partial_i \partial_j p - \partial_j \partial_i p = 0$. Also, observe that $\mathsf{Div}(pX^{ij}) = X_{ij} \cdot \nabla p = 0$. From $\mathsf{Div}(pX^{ij}) = 0$, we can infer that the $(q_t)_*\mathbb{P} = \mathbb{P}$ and from $\mathsf{Div}(X^{ij}) = 0$, Φ_t is a volume preserving diffeomorphism. Consider the following autoencoder where the decoder is $\tilde{f}_t = f \circ q_t$ and the encoder is $\tilde{g}_t = q_t^{-1} \circ f^{-1}$. Observe that these autoencoders achieve perfect reconstruction. The encoder fails at identifying the underlying latents as the estimated latents are related to the true latents by the map $\Phi_t^{-1}(\cdot)$. Thus we have seen that even if the learner knows the true density p_Z, there exists a family of autoencoders that cannot achieve identification.

Suppose $Z = [Z_1, Z_2]$. Consider that p_Z is a density defined over the unit square centered at origin. In Figure 14, we show one such density. Suppose we are interested in separating the points in the four quadrants, where each quadrant provides a distinct quantization for all the values of z assumed in it. We consider a family of densities p_Z, whose isolines crosses $z_1 = 0$ at least once. We further also assume that these densities p_Z are differentiable over the support and have a bounded gradient. The isoline of this density crosses $z_1 = 0$. Consider two points shown in pink and blue colors in Figure 14. At $t = 0$, the pink point is to the right of $z_1 = 0$ and the blue point is to the left of $z_1 = 0$. Under the flow defined in equation (4), we would argue that after some time τ has elapsed, the pink point moves to the left of $z_1 = 0$, while the blue point is still to the left of $z_1 = 0$. Under the map $\Phi_\tau(z) = (q^1_\tau(z), q^2_\tau(z))$, the two points are both to the left of $z_1 = 0$. If all the points $\Phi^1_\tau(z) < 0$ are associated with the same quantization, then the pink point and the blue point are associated with same quantization, while their true quantization is different. We provide further details on the construction. We can assume that the pink point at $t = 0$ is very close to $z_1 = 0$ but on the right of $z_1 = 0$. Further, we assume that the magnitude of the flow along the negative z_1 direction (which is the vector tangent to the isoline) in the neighborhood of the pink point at $t = 0$ is bounded below by at least ϵ_1. As a result, the point moves at least $\tau\epsilon$ distance along z_1 in time τ. We can assume that the pink point started with $z_1 < \epsilon\tau$ and thus it crosses to the left of $z_1 = 0$ after τ amount of time has elapsed. At the same time, consider the blue point to the left of $z_1 = 0$ at time $t = 0$. We assume that the flow along the z_1 direction in the ρ radius neighborhood of the blue point at $t = 0$ points from right to left, i.e., it is in the negative z_1 direction. We choose τ to be sufficiently

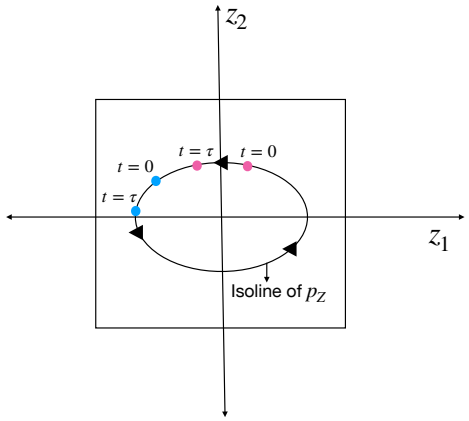


Figure 14: Two dimensional illustration to explain why isolines cannot cross the axis.

small, $\tau < \rho/\gamma$, where γ is the largest value that the velocity under the flow can take (this follows from the assumption that the density has a bounded gradient). Under this constraint, the blue point stays to the left of $z_1 = 0$ after τ amount of time has elapsed.

The above argument explains that if the isolines of p_Z cross the grids, then quantized identification is not possible even if we know p_Z. As a result, we need to focus on the densities p_Z whose isolines are restricted to each of the four quadrants. Consider the family of densities p_Z with axis-aligned discontinuities. Suppose the density in each of the quadrants is continuous, then for this class of densities the isolines cannot cross any of the axis $z_1 = 0$ and $z_2 = 0$. For this class of densities, our theory established quantized factor identification guarantees even without requiring knowledge of p_Z. Could there be other densities beyond discontinuous densities with axis-aligned landmarks that permit quantized identification? This is a fairly non-trivial and important question left for future work.