# The Dimensions of Dimensionality

Brett D. Roads

Department of Experimental Psychology, University College London, London, United Kingdom, WC1E

Bradley C. Love

Department of Experimental Psychology, University College London, London, United Kingdom, WC1E

Correspondence: b.roads@ucl.ac.uk (B.D. Roads)

**Keywords:** dimensions, embedding, manifold, latent representations, dimensionality

**Abstract:** Cognitive scientists often infer multi-dimensional representations from data. Whether the data involve text, neuroimaging, neural networks, or human judgments, researchers frequently infer and analyze latent representational spaces (i.e., embeddings). However, the properties of a latent representation (e.g., prediction performance, interpretability, compactness) depend on the inference procedure, which can vary widely across endeavors. For example, dimensions are not always globally interpretable and the dimensionality of different embeddings may not be readily comparable. Moreover, the dichotomy between multidimensional spaces and purportedly richer representational formats, such as graph representations, is misleading. We review what the different notions of dimension in cognitive science imply for how these latent representations should be used and interpreted.

# Glossary

- Observable Data: Data associated with a directly observable measurement (e.g., a photodiode measuring photon wavelength).

- Input Data: Data that is input into an embedding algorithm. The input data may be directly observed or the outcome of a previous processing step. Input data does not have to be multidimensional, for example it may be text or graph-like.

- Embedding or Encoding Algorithm: An algorithm that trains a mathematical entity, sometimes called an encoder, using input data and (optionally) target data.

- Encoder: A mathematical entity that maps from input space to a latent space.

- Latent or Embedding Space: The learned multidimensional space that is a transformation of the input space. Typically, embedded data captures some notion of proximity, i.e., similar items from the input data tend to be neighbors in the embedding. The specific notion of proximity is determined by the particular algorithm used to infer the embedding.

- Dimensionality: The number of dimensions present in a given multidimensional representation, which is a consequence of both the input data and the embedding algorithm.

- Manifold: A relatively contiguous lower-dimensional space that exists within a higher-dimensional space.

# Inferring the unobservable

Formal models of cognition are a cornerstone of cognitive science. Broadly speaking, formal models use various mathematical entities—such as vector spaces and functions—to explicitly describe *mental representations* (the content) and *mental processes* (how mental content is transformed). When a cognitive model adequately mirrors a biological target, the model provides a tantalizing glimpse of the representations and processes used by the brain. Modeled representations reveal how people experience the world and open the door to answering age-old questions, such as "Do two people experience the color red the same way?". But modeled representations can do so much more, they can reveal what people notice and what they miss. For example, a dermatologist may quickly recognize that a skin lesion is malignant while a novice remains oblivious to the diagnostic features. In other words, modeled representations reveal the boundaries of thought; the frontiers of one's experienced reality. Unfortunately mental representations cannot be directly observed, they must be inferred. And the inference process opens Pandora's box.

Since mental representations are not directly observable, the inferred mathematical entities are often referred to as *hidden representations*, *latent representations*, or *embeddings*. To infer latent representations, researchers leverage **observable data** (see Glossary) from a variety of sources such as human judgments, neuroimaging data, image datasets, large text corpora, and neural network models. Across these diverse sources of data, the general objective remains the same; discover *informative* latent dimensions. For example, it is not the pixels in a photograph that are interesting, but the increasingly abstract latent features (such as edges, corners, and faces) that empower the brain to make sense of the world. In other words, observable data are merely a stepping stone towards revealing the statistical regularities that a brain exploits.

A rich space of techniques exists for extracting insightful latent dimensions, loosely referred to as **encoding algorithms** or **embedding algorithms**. Widely used examples include Principal Component Analysis (PCA) [1, 2] and t-SNE [3], but a bustling zoo of familiar and exotic algorithms exists for discovering latent dimensions [1, 2, 4, 5, 3, 6–8]. Although all embedding algorithms share the same general objective, there are meaningful differences. On the input side, different algorithms place restrictions on the type of data they can ingest. On the output side, different algorithms prioritize different **latent space** properties—such as prediction performance, transformation type, interpretability, and compactness. Collectively, these differences create a space of trade-offs, where researchers must decide which properties best suit their goals. Seemingly minor algorithmic differences can have unintuitive consequences for the inferred latent space. For example, the latent space dimensions may have limited interpretability. Or the **dimensionality** of the latent space may be unrelated to the number of so-called "true" dimensions of the task. Or structurally dissimilar latent spaces may fit the **input data** equally well. The remainder of this article expands on key considerations when inferring latent spaces. Before diving in, it is worth pausing for a moment to unpack the metamorphosis that occurs during an embedding transformation.

## Dimensions before and after

A dimension is a basic building block that contributes to a multidimensional structure. But the meaning of a dimension changes when mapping from dimensions defining an observable space to those defining a latent space. Observable data is fundamentally anchored in reality; it is as close as we can get to an objective measure of something that physically exists. In the case of multidimensional observable data, each dimension quantifies how a specific observable feature varies, such as the wavelength of light absorbed by a single photodiode.

After observable data are embedded, the meaning of a dimension is altered in two fundamental ways. First, the new latent dimensions are not guaranteed to correspond to something globally interpretable. In other words, the new dimension may not have an easy to articulate label since each latent dimension may use a sprinkling of all observable dimensions. We return to the issue of interpretable dimensions later.

The second fundamental difference is that latent dimensions do not necessarily exist in the real world. Latent dimensions are unobserved variables that can be estimated from input data. When someone looks at a photograph of a bird, the latent features burst into awareness so rapidly that there is an illusion that the features always existed *out there.* But photons do not carry metadata stating "the last thing I bounced off of was an orange beak". Latent visual features like an orange beak do not actually exist in the visual world—they must be inferred and rendered by the perceptual machinery of our minds. You might be thinking, of course the bird beak exists. While that is true, the bird exists in terms of *physical* atoms. The *visual* features are inferred by our mind and at least one processing step removed from reality. This distinction is made apparent by converting the pixels of a photograph into a table of numbers. Reformatted this way, our existing perceptual machinery is completely derailed and the latent dimensions fail to materialize in our mind's eye.

At times, it is easy to think nothing profound has occurred when mapping from observable dimensions to latent dimensions, but *good* latent dimensions are the heart of cognitive models. Despite the unquestionable utility of latent dimensions, “latent” is still a polite way to refer to something made-up, albeit in a data-driven way. Given the data-driven aspect of embeddings, it is important to recognize that an new set of input data could drastically change the inferred latent space. There is a balance to be made between researchers sharing out imaginary (but promising) dimensions and reminding fellow researchers to hedge their bets. In the absence of extraordinary evidence, the research community should remain cautious of designating a particular latent space as having the final say. At any time, a new embedding approach could be invented or additional data could become available that shakes up the status quo.

# Engineering desirable latent dimensions

Observable data by itself, provides limited insight and must be processed using algorithms that reveal valuable informative dimensions. It may be tempting to believe that the inferred latent dimensions are the objectively “true” dimensions. However, it is important to recognize that different algorithms optimize different properties; meaning that there is no single latent space that is objectively correct. Rather, the optimal latent space is the one the best suites the research question. For example, both the Big Five Inventory[9, 10] and HEXACO [11] provide low-dimensional latent spaces for describing personality traits, but HEXACO may make better predictions for moral and ethical behavior. Likewise, inferred latent dimensions are highly dependent on the input data. When the input data is primarily derived from undergraduates in western universities, the latent dimensions may not be representative of the wider world [12].

When selecting an embedding algorithm there are five key considerations. On the input side, (1) one must select an algorithm that accommodates the structure of the input data. On the output side, different algorithms prioritize different latent space properties: (2) prediction performance, (3) transformation type, (4) interpretability, and (5) compactness. A small preview of embedding algorithms and their properties are summarized in Table 1. For a brief discussion of practical considerations see Text Box 2. These properties often trade-off against one another, forcing researchers to prioritize based on their objectives. Each of these considerations is unpacked in more detail below.

## Input source and structure

Observable data can be obtained from remarkably different sources. Intuitive examples include measurements of sensory modalities, such as as photon wavelengths, audio frequencies, and the outputs of gas chromatography-olfactometry. Various techniques also yield observable data beyond everyday sensory information. MRI sensors detect the energy released from protons, yielding multidimensional voxel data. EEG measures electrical activity, where each electrode channel yields a separate dimension.

Embedding algorithms can directly ingest observable data, but more generally ingest **input data** that is derived from observable data, but has already undergone an arbitrary

Table 1: Properties of a handful of latent space algorithms.

| Method | Input Structure | Transformation |
|---|---|---|
| PCA | multidimensional | linear |
| ICA | multidimensional | linear |
| Factor analysis | multidimensional | linear |
| Fourier transform | multidimensional | linear |
| Linear Discriminant Analysis | multidimensional | linear |
| Canonical Correlation Analysis CCA | multidimensional | linear |
| Kernel PCA | multidimensional | nonlinear |
| Isomap | multidimensional | nonlinear |
| Locally linear embedding (LLE) | multidimensional | nonlinear |
| Hessian LLE | multidimensional | nonlinear |
| Laplacian eigenmaps | multidimensional | nonlinear |
| Generalized discriminant analysis (GDA) | multidimensional | nonlinear |
| Autoencoders | multidimensional | nonlinear |
| Nonnegative Matrix Factorization (NMF) | multidimensional | nonlinear |
| t-SNE | multidimensional | nonlinear |
| t-STE | inequalities | nonlinear |
| UMAP | inequalities | nonlinear |
| Poincaré | inequalities, graphs | nonlinear |
| GloVe | text corpora | nonlinear |
| Word2Vec | text corpora | nonlinear |

number of transformations. For example, raw photodiode readings are first converted into discrete pixel values via low-level photo-processing steps. Likewise, fMRI signals are first converted into calibrated voxel data via sophisticated processing pipelines like fMRIPrep [13]. Embedding steps can also be chained to create an arbitrary sequence of data transformations where the output of one embedding step is passed as input to the next embedding step; as is the case in deep neural networks.

So far we have outlined how input data varies along two axes: (1) variety in source and (2) variety in the extent of prior processing. A third axis concerns the *structural form* of input data. For example, the input data may be arranged as a table of numbers (i.e., multidimensional data), a set of ordinal relationships (e.g., human preference judgments), a graph of nodes and edges, or a large body of text. Historically, embedding algorithms have focused on ingesting multidimensional data, such as in the case of PCA [1, 2], ICA [4], autoencoders [7, 8], and t-SNE [3]. But embedding algorithms can *embed* arbitrary input data in a multidimensional space.

In cognitive science, a common source of data comes from collecting human similarity judgments (e.g., stimulus Q is more like stimulus A than stimulus B). A diverse family of embedding algorithms exist that transform ordinal similarity judgments into a multidimensional representation where each stimulus is represented as a point in the latent space [14–27]. Many of these algorithms do not have formal names, but fall into the general family of *multidimensional scaling algorithms* or *psychological embeddings*. The inferred latent dimensions formalize the notion of psychological distance; where similar items occur close together (Figure 1e). Analogous embedding algorithms also exist for categorization confusion matrices, pairwise ratings [28, 29], odd-one-out judgments [30], arrangement data [31, 32], and non-human paradigms [33, 34]. Embedding algorithms can also be guided by asking participants to rate the degree that a particular stimulus (image of a cat) exhibits a particular feature ("is stealthy") [35, 36, 29]. In all of these efforts, the similarity function chosen to model behavior may be pragmatic or reflect theoretical assumptions, such as how the brain represents and processes information [37].

Embedding algorithms can also ingest text corpora in order to produce multidimensional word embeddings [38, 39] and sentence embeddings [40]. Perhaps easiest to understand are word embedding algorithms that leverage the co-occurrence statistics of words, which counts how often a pair of words co-occur within a specified window (Figure 1f). Words that co-occur in similar contexts likely have similar meanings. For example, pet animal words are likely to co-occur with the phrase "water bowl". Word embeddings have also been extended so that every word is represented as a multivariate Gaussian distribution in the multidimensional space [41]. In such a space, the covariance matrix of the Gaussian can capture properties like lexical entailment. For example, the concept "mammal" may exhibit a large Gaussian that overlaps the concepts "bear", "otter", and "whale". Modern multidimensional word embeddings have been shown to capture linguistic hierarchical structure [42], further demonstrating the breadth of possible meanings for multidimensional spaces.

The notion of latent dimensions can even capture graph structure, where connected nodes are embedded as nearby points (Figure 1d). Graph embeddings deserve special mention because hierarchical graphs have long been argued as a case that multidimensional spaces struggle to adequately model [43–45]. Modern embedding algorithms have demonstrated that graphs can be embedded in multidimensional spaces [46–48]. Hierarchical graphs can be embedded in multidimensional spaces by exploiting properties of hyperbolic spaces [49, 50], which tend to require fewer dimensions relative to their Euclidean counterparts when modeling hierarchical data [49]. Some behavioral data, such as smells, are well-described by a hyperbolic space [51]. Graph embedding algorithms highlight how the distinction between graph data and multidimensional data can be superficial. Sometimes it is possible to generate topological signatures to assess whether data are better described by the curvature in a Euclidean or hyperbolic latent space [52], which have been used to argue spatial maps are hyperbolic [53].

## Prediction performance

Perhaps the most important property of an embedding algorithm is the preservation of information when mapping from input space to latent space. A good latent space will retain the key statistical regularities present in the input data while discarding noise. The appropriate metric for assessing preservation of information will vary by application, but can be loosely described as the *prediction performance* of the latent space. For example, after applying PCA one could measure the total variance explained by the latent dimensions. Likewise, if the latent representation is used in a downstream image classifier, then the embedding algorithm could be scored on its ability to embed a test image among neighbors of the same category. Prediction performance can be expounded further by considering whether an approach permits out-of-sample predictions and estimates of uncertainty.

Prediction performance may vary for trained items (i.e., within-sample) versus test items (i.e., out-of-sample). Out-of-sample predictions are important because they enable an **encoder** to generalize to novel situations. For example, a mammogram encoder will have limited practical value if it cannot embed X-ray images obtained from new patients. Interestingly, many algorithms cannot perform out-of-sample generalization, although much work has been done to extend popular approaches [54, 55]. In contrast, artificial neural networks naturally make predictions for novel inputs. Out-of-sample generalization may suffer from overfitting, which can be framed as a separability versus generalization trade-off [56] or an efficient versus robust trade-off [57].

A second factor in prediction performance is the ability to generate uncertainty estimates. Algorithms can learn point estimates or distributions for each embedded sample. Learning an embedding with uncertainty estimates has the advantage of describing both the most likely location (e.g., mode) of the embedding point and how confident we can be of the embedded location. Various techniques are available to learn uncertainty embeddings, such as variational inference autoencoders [8], variational psychological embeddings [23] and epistemic neural networks [58]. Uncertainty information can help inform downstream analysis, such as identifying latent space regions that exhibit individual differences, estimating the significance of a treatment condition, or computing expected information gain within a active learning paradigm [23, 59].

## Transformation type

The type of transformation has a large impact on the final inferred latent space. Broadly speaking, the type of transformation can be characterized by the linearity of the embedding transformation and the change in dimensionality. A transformation of the input space is typically (but not always) performed with the additional goal of discovering a *lower* dimensional latent space. A lower dimensional latent space is typically desirable because it transforms a large number of (likely) uninformative dimensions into a smaller number of informative dimensions. For example, face space is a low dimensional space that can be used to describe the perceived similarity between human faces and can help explain face distinctiveness [60]. In contrast, an embedding transformation can also be used to discover a *larger* dimensional latent space, such as when ICA is applied to image pixels to discover a rich set of latent dimensions that look remarkably like V1 receptor fields [61].

When restricted to linear transformations, the new dimensions embody relatively simple changes (Figure 1a-b). Popular methods of linear transformations include PCA [1, 2], ICA [4], Fourier transform, factor analysis [62], and tensor-based dimensionality reduction [63, 64]. In contrast, a nonlinear transformation—such as t-SNE [3]—allows the new dimensions to describe a drastically different space, potentially twisting and turning through the original space (Figure 1c).

The consequences of dimensionality reduction are different for linear and nonlinear approaches. For linear approaches, dimensionality reduction may involve identifying a set of latent dimensions that explain the most variance and excluding any remaining latent dimensions (Figure 1a). For example, if two input dimensions are highly correlated then one of the dimensions can be dropped. In contrast, nonlinear dimensionality reduction algorithms often assume that data are not distributed in a uniform way, but that the data points are distributed in relatively contiguous sheets (i.e., **manifolds**) that twist and turn in a larger multidimensional space. When the manifold assumption is true, the empty parts of the original space can be "squeezed" or "flattened" out (see Text Box 1 for an example). This assumption is known as the "manifold hypothesis". Widely used nonlinear dimensionality reduction algorithms include ISOMAP [5], t-SNE [3], UMAP [6], and various flavors of autoencoders [7, 8]. Other nonlinear dimensionality reduction algorithms include kernel methods [65], local linear embedding [66], Laplacian eigenmaps [67], Hessian eigenmaps [68], and local tangent space alignment [69].

Cognitive scientists have developed a diverse set of application-specific dimensionality reduction methods. Neuroscience has been particularly prolific in developing dimensionality reduction algorithms that can ingest neural population activity [70]. Neuroscience specific methods range from diffusion embedding for connectivity data [46], delayed latents across groups (DLAG) for neurophysiological recordings [71], and calcium imaging linear dynamical system (CILDS) [72]. The concept of low-dimensional manifolds has invigorated discussions around the brain's latent representations, proving fruitful for studying motor [73] and spatial representations [53].

## Interpretability

Cognitive scientist are often interested in uncovering psychologically interpretable dimensions in order to advance mechanistic hypotheses. For example, latent dimensions can define a motor space with separate preparation and execution dimensions [73]. While the interpretation of a observable dimension is self-evident, the interpretation of a latent dimension is not straightforward. The specifics of the embedding algorithm—such as linear or nonlinear latent dimensions—influence the manner of interpretation.

There are two forms of interpretability. In its strong form, interpretability means that *each* dimension has a clear interpretation independent of the other dimensions. For example, the second dimension of a latent space may correspond to "wingspan of a bird" where lower values mean smaller wingspan. The strong form can be framed as *global interpretability*—the interpretation holds for the entire dimension regardless of where you are on the dimension. This strong form of interpretability is often referred to as a *disentangled representation* in machine learning [74]. In contrast, a dimension may only have *local interpretability*—the interpretation of the dimension changes as you gradually move along the dimension (Figure 2c). Given sufficient input data that is not pure noise, an embedding should at least have local interpretability. In practice, an embedding may not appear to exhibit local interpretability if the input data contains too few samples. For both global and local interpretability, although the dimensions capture a clear statistical regularity, it may be difficult to articulate a description of the dimension using natural language.

In our estimation, when cognitive scientists discuss dimensions they are often referring to globally interpretable dimensions, perhaps because such dimensions are the most intuitive and the simple stimuli of early methods yield such spaces. The colloquial prominence of globally interpretable dimensions is likely encouraged by cognitive models that deploy dimension-wide attention to expand or shrink the extent of a dimension, thus altering perceived similarity [75–77]. Relatedly, globally interpretable latent dimensions are a highly desirable outcome since such representations provide actionable insight, such as helping researchers understand how representation spaces in the brain change as a function of learning [78–81]. Interpretable latent dimensions could be used to design more efficient training programs. For example, if the latent dimensions reveal a diagnostic dimension of malignant skin lesions, then training can be structured to emphasize learning the diagnostic dimension.

If globally interpretable dimensions are a high research priority, this property should be deliberately balanced with other properties. To bias the inferred latent space towards globally interpretable dimensions, one option is to use algorithms that employ non-negativity constraints that force the discovery of sparse, part-like representations [30, 20, 82]; although this will likely come at the cost of needing a latent space with substantially more dimensions. By employing non-negative constraints, each dimension is biased to code for part-based features, promoting compositional representations that tend to be easy to articulate. Similarly, demixed PCA is capable of exposing the dependence of the neural representation on task parameters such as stimuli, decisions, or rewards [83, 84].

## Compactness

In light of the properties introduced above, one can see why it is problematic to make absolute claims about the number of "true" latent dimensions. For example, one could trade a lower dimensional, locally interpretable space for a method that yields a higher dimensional, globally interpretable space. Although both solutions may fit the input data equally well, they will differ in the number of dimensions recovered. Going the other direction, a researcher could employ non-linear transformations or hyperbolic spaces in order to obtain a compact lower dimensional representation. When work indicates that the measured dimensionality of the neural representations in PFC is high [85], that neural codes are confined to low-dimensional latent spaces [70, 84], or that neural codes actually exist in high-dimensional spaces [86], it is important that these results are also qualified by employed embedding technique and the corresponding properties being prioritized.

A potential improvement over reporting dimensionality, is to report the *intrinsic* dimensionality, which can be thought of as the smallest dimensionality possible at the expense of all other properties. However, methods for computing intrinsic dimensionality can still be sensitive to factors like the distribution of points and curvature of the latent space, so care must be taken to use modern methods that are robust to these complications [87, 88].

Regardless of reporting dimensionality or intrinsic dimensionality, it is our view that researchers should make clear which latent space properties are being prioritized (e.g., prediction performance, globally interpretable dimensions, compactness). Dimensionality by itself is not a good indicator of the amount of information in a representational space. Rather, one should aim to make dimensionality comparisons between representations that are derived with comparable assumptions [79]. High-quality work that includes analyses of the functional or intrinsic dimensionality of a dataset [86, 89] is still bounded by the assumptions baked into intrinsic dimensionality computations.

## Selecting and comparing latent spaces

If one is seeking a one-size-fits all embedding solution, the trade-offs introduced above should make it clear that such a solution does not exist. The natural course of research will inevitably lead to a need to select the latent space(s) that best meet the research objective. Model selection can be performed by directly comparing candidate embeddings spaces or indirectly comparing embeddings via performance on a downstream task.

A direct comparison can identify how two latent spaces agree and diverge, which in turn suggests information that is shared or distinct between the two sources of input data. For example, one may want to know whether the latent representation of fMRI data from a particular brain region is capturing high-level semantic information (such as “this item is edible”) versus lower-level feature information (“this item is yellow”). One way to address this question is to compare the latent representation of fMRI data to latent representations based on category membership (limes and tennis balls will be in distinct clusters occurring far apart) versus latent representations based on low-level features (limes and tennis balls will be intermixed because they are both round and green) [90]. Likewise, one may want to compare the latent represent of fMRI data from a particular brain region to the latent representation of different layers of a deep neural network in order to find which parts of a neural network best correspond to specific regions of a brain [91]. Many different types of representation comparisons are popular in cognitive science, such as brain versus behavior, model versus brain, model versus model [92], and language versus language. For example, the function of a brain region can be clarified by finding a correspondence between a latent space derived from a region’s brain activity and a latent space derived from a cognitive model fit to behavior [93]. All of these comparison enable researchers to quantify differences and highlight deficiencies in a candidate embedding.

Latent spaces are in agreement if they display second-order isomorphism [94]. For example, the image of a crow is the nearest neighbor of the image of a raven in both latent spaces. Second-order isomorphism does not require the coordinate values or the dimensionality of the two spaces to match. Instead, the notion of match is more abstract—the patterns between the points in the two spaces need to display the same relations. For example, consider a set of points randomly arranged in a two-dimensional space. If one made a copy of this space and rotated it by 90 degrees, the new space would have a completely different set of coordinate values, suggesting that the two representations are different. If instead one considered the relationships between the points—such as pairwise distances—one would conclude that the two spaces describe an identical set of relationships. In this simple example, the two spaces are perfectly isomorphic, but in practice matches are imperfect and evaluating alignment between two spaces can be challenging.

The optimal method for comparing representations is an ongoing research problem. Numerous techniques exist and each has limitations and built-in assumptions. Popular techniques for comparing representations include representational similarity analysis (RSA) [95, 90] and canonical correlation analysis (CCA) [96–98]. Briefly, RSA is a method for comparing two representations that assesses the correlation between the implied pairwise similarity matrices. CCA is a method that compares two representations by finding a pair of latent variables (one for each domain) that are maximally correlated. While RSA and CCA remain popular, their limitations have spurred researchers to develop more robust methods [99, 100]. Adjustments to RSA include the similarity metric Centered Kernel Alignment (also known as the RV coefficient) [101–103], unbiased CKA [104], feature-reweighted RSA [105], and extensions for handling noise [106]. Numerous CCA variants have been introduced to make the approach more widely applicable, such as probabilistic CCA [107, 108], kernel CCA [109], deep CCA [110], sparse CCA [111], and projection weighted CCA [92]. Techniques also exist beyond RSA and CCA, such as pattern component modeling [112].

One practical strategy for selecting between different latent spaces is to assess performance on a downstream task. For example, one could evaluate the ability of deep neural network embeddings to correctly predict different aspects of human behavior, such as similarity judgments [18, 23], categorization performance [113, 114], categorization performance using degraded images [115] and trial-by-trial categorization error consistency [116]. The advantage of this strategy is that it allows comparing latent spaces with very different assumptions since the downstream task can be agnostic to the particular embedding details. The drawback of this approach is two-fold: researchers must define how the embedded coordinates are mapped to observable behavior and the target behavioral data (e.g., categorization performance) may not be available. In practice, one can evaluate multiple methods for assessing correspondence and choose the one that performs best on some gold standard. Following machine learning best practice, one suggestion is to use holdout data, or simulated data to select the most appropriate procedure.

## Concluding remarks

Embedding algorithms provide an exciting and powerful technique for understanding the content of mental representations. Embedding arbitrary input data is a key component of cognitive science research with many open avenues for research (see Outstanding Questions). When studying and communicating results it is important to realize that different algorithms bestow different interpretations on the recovered dimensions. Multidimensional representations are extremely flexible: capable of condensing high-dimensional representations down to low-dimensional spaces, representing part-like features with high interpretability, and capturing graph-like relationships. The extreme flexibility of multidimensional spaces means that superficially different representations can capture underlying structure equally well. As a consequence, gross measures like dimensionality convey little on their own; more dimensions does not mean more information. Instead, researchers need to make controlled comparisons paying attention to the strengths and weaknesses of different comparison methods. Defaulting to a single comparison method like intrinsic dimensionality, CCA, or RSA risks overlooking novel insight. While this focuses on comparisons between latent spaces for purposes of data analysis and theory evaluation, one exciting possibility is that biological systems also balance these trade-offs (see Text Box 3) and learning itself may proceed by comparing and aligning different latent spaces in an unsupervised fashion [117].

## Acknowledgments

This work was supported by an ESRC Grant ES/W007347/1 and a Royal Society Wolfson Fellowship 183029 to B.C.L.

## Declaration of interests

BDR is a part-time contract worker at Magnit Global @ Meta.

## Elements

**Figure 1 Caption:** Examples of latent dimensions derived from six different datasets. Each sub-panel depicts the input data and any input dimensions (gray), followed by the latent dimensions (red) that best capture the structure of the data. **(a)** Latent dimensions are linear and orthogonal to one another (e.g., PCA). **(b)** Latent dimensions are linear, but not necessarily orthogonal to one another (e.g., ICA). **(c)** Latent dimensions can be nonlinear, curving through the original space (e.g., ISOMAP, t-SNE, UMAP). **(d)** Latent dimensions are derived from a hierarchical graph and embedded in a hyperbolic space (e.g.,

Poincaré embedding). **(e)** Latent dimensions are inferred from ordinal similarity relations. For example $s_1$: $s_2 > s_3$ means that given stimulus $s_1$, participants judge $s_2$ to be more similar than $s_3$. **(f)** Latent dimensions are inferred from a text corpora.

**Figure 2 Caption:** Example demonstrating the difference between global and local interpretability of multidimensional spaces. **(a)** input dimensions (columns) of a set of input data composed of nine concepts (rows). A darker cell means more of a particular feature. **(b)** A two-dimensional projection of the first two input dimensions where each dimension is globally interpretable. **(c)** An analogous non-linear embedding of the input data where the latent dimensions are locally, but not globally interpretable.

**Text Box 1: Manifold example** As a simple example of a twisty manifold, consider a set of images that are each 100×100 pixels, where each pixel can be either black or white. Each image depicts a single black circle on a white background. All images are unique since the circles vary in both size and location. In its input form, each image has 10,000 dimensions (one dimension for each pixel). With respect to input space, each image can be thought of as a point in the 10,000 dimensional space. Even with many unique images, vast regions of the 10,000 dimensional space will be unoccupied. For example, you will never encounter an image with an isolated black pixel, so all coordinates corresponding to an image with isolated black pixel will be unused. Occupied parts of the space will tend to be clumpy—circles that have a similar radius or slightly different center will occur close together in pixel space. The vast unused regions of pixel space suggest that a lower dimensional latent space is possible. By construction, we know this to be true. The data associated with these images can be redescribed using three latent dimensions: an x- and y-coordinate describing the center of the circle and the radius of the circle. The new latent space is more concise and the underlying structure of the stimuli is made clearer.

**Text Box 2: Practical considerations** In addition to key theoretical considerations, practical considerations must also be taken into account when selecting an embedding algorithm. The impact of computational efficiency varies by application, but typically becomes more important as the problem size grows. For small problems on the order of a few hundred data points, popular embedding algorithms easily find solutions within a few seconds on widely available hardware and most computational inefficiencies can be ignored. But as problems grow, it becomes increasingly difficult to ignore computational inefficiencies because poor design choices can preclude discovering solutions within a reasonable timeframe. Some algorithms will be more or less prone to getting stuck in local optima and finding degenerate solutions. Different algorithms will also have varying memory and storage requirements. For example, non-negative embedding algorithms typically require substantially more dimensions—thus more memory and storage—than an equally accurate unconstrained embedding algorithm. In this case, there is a substantial trade-off between interpretability and model size. In a similar vein, estimating uncertainty using Markov chain Monte Carlo methods will scale poorly compared to variational inference [118].

**Text Box 3: Trade-offs in biological systems** Biological systems, like researchers, also aim to balance potentially competing demands. The balance can shift depending on the task and available resources. In some cases, interpretability may be of primary importance whereas in other cases compactness may be paramount. Almost all of the points above also apply to an organism's point of view. For example, brain regions need to communicate with one another and eventually affect behavior. This "inside view" is sometimes neglected but should be adopted to understand the function of biological systems.

The nature of the neural computation may determine what is the best representational format. For example, when information is passed from one region to another region via biological neural networks akin to a linear transformation then globally interpretable dimensions may offer advantages. In this case, a linear transformation can more readily extract the relevant information than when the same information is embedded in a lower-dimensional manifold where dimensions have no obvious interpretation. Moreover, globally interpretable dimensions may offer communicative benefits as they can readily align with language (e.g., the dimension of size easily maps to codified relations in language such as "bigger than").

In other cases, the opposite may be true and the brain may choose not to decorrelate different dimensions in order to promote redundant communication across brain areas [119] or a brain region may favor cells that have a mixed selectivity to boost coding capacity [89]. Likewise, low-dimensional manifolds with local interpretability may better support inter-region communication when the readout network functions as a tuned receptive field, as in artificial radial basis networks. For example, the sparse coding of cells that selectively respond to a particular celebrity [120] may be repackaged into a low-dimensional representation to facilitate inter-region communication.

Latent representations may also exhibit varying degrees of specialized compartmentalization. At one extreme, a region may create a vast unified latent space, such as recently proposed for category-selective visual regions [121]. At the other extreme, rather than extracting a single shared latent space, the brain may find it worthwhile to maintain a set of (nearly) orthogonal subspaces [122], as in the case of whisker contacts in rodents [123]. Resource costs in terms of required cells, metabolism, and wiring may lead the brain to favor (when possible) more compact solutions that lack global interpretability, but invest in more expensive solutions when it pays off.

Given the different trade-offs associated with latent space properties, it is plausible that different brain regions emphasize different properties. The success of neuroscientists using imaging techniques with limited spatial and temporal resolution in uncovering embedding spaces suggests that in some cases neural representations are somewhat smooth and regular [124]. The brain itself is not uniform in its circuitry and function, which means that globally interpretable dimensions may be less likely in certain regions, such as prefrontal cortex [125]. Even within a region such as prefrontal cortex, the dimensionality found may be quite low [126] or high [127], perhaps due to task differences. For example, over-training on a task may lead to fine-grain, essentially overfit, representations. Even within the same region of visual cortex with the same stimuli, minor task differences in how many aspects of a stimulus is relevant for a decision can affect dimensionality estimates [79]. Activity in some brain regions indicates representations live in a hyperbolic latent space [128], such as spatial representation in the CA1 region of the rodent hippocampus [53] and early visual cortex [129]. One possibility is that people rely on multiple embedding spaces with the relative importance of each varying with context [130].

# References

[1] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572, 1901. doi: 10.1080/14786440109462720.

[2] Diogo Manoel, Melanie Makhlouf, Charles J. Arayata, Abbirami Sathappan, Sahar Da'as, Doua Abdelrahman, Senthil Selvaraj, Reem Hasnah, Joel D. Mainland, Richard C. Gerkin, and Luis R. Saraiva. Deconstructing the mouse olfactory percept through an ethological atlas. *Current Biology*, 31(13):2809–2818.e3, 2021. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2021.04.020. URL `https://www.sciencedirect.com/science/article/pii/S0960982221005339`.

[3] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[4] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994. doi: https://doi.org/10.1016/0165-1684(94)90029-9.

[5] Vin Silva and Joshua Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. URL `https://proceedings.neurips.cc/paper/2002/file/5d6646aad9bcc0be55b2c82f69750387-Pap`

[6] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861.

[7] Geoffrey E. Hinton and Richard S. Zemel. Autoencoders, minimum description length and helmholtz free energy. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, page 3–10, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[8] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[9] Robert R McCrae and Antonio Terracciano. Universal features of personality traits from the observer's perspective: data from 50 cultures. *Journal of personality and social psychology*, 88(3):547, 2005.

[10] David P Schmitt, Jüri Allik, Robert R McCrae, and Verónica Benet-Martínez. The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of cross-cultural psychology*, 38(2):173–212, 2007.

[11] Michael C Ashton, Kibeom Lee, Marco Perugini, Piotr Szarota, Reinout E De Vries, Lisa Di Blas, Kathleen Boies, and Boele De Raad. A six-factor structure of personality-descriptive adjectives: solutions from psycholexical studies in seven languages. *Journal of personality and social psychology*, 86(2):356, 2004.

[12] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3):61–83, 2010. doi: 10.1017/S0140525X0999152X.

[13] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik,

Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit S. Ghosh, Jessey Wright, Joke Durnez, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fmriprep: a robust preprocessing pipeline for functional mri. *Nature Methods*, 16(1):111–116, 2019. doi: https://doi.org/10.1038/s41592-018-0235-4.

[14] Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert Lanckriet, David Kriegman, and Serge Belongie. Generalized non-metric multidimensional scaling. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 11–18, San Juan, Puerto Rico, 3 2007. PMLR.

[15] Siavash Haghiri, Felix A. Wichmann, and Ulrike von Luxburg. Estimation of perceptual scales using ordinal embedding. *Journal of Vision*, 20(9):14–14, 09 2020. doi: 10.1167/jov.20.9.14.

[16] Emilie L. Josephs, Martin N. Hebart, and Talia Konkle. Dimensions underlying human understanding of the reachable world. *Cognition*, 234:105368, 2023. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2023.105368.

[17] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[18] David-Elias Künstle, Ulrike von Luxburg, and Felix A. Wichmann. Estimating the perceived dimension of psychophysical stimuli using triplet accuracy and hypothesis test-

ing. *Journal of Vision*, 22(13):5–5, 12 2022. ISSN 1534-7362. doi: 10.1167/jov.22.13.5. URL `https://doi.org/10.1167/jov.22.13.5`.

[19] L. van der Maaten and K. Weinberger. Stochastic triplet embedding. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6, 9 2012. doi: 10.1109/MLSP.2012.6349720.

[20] Lukas Muttenthaler, Charles Yang Zheng, Patrick McClure, Robert A. Vandermeulen, Martin N Hebart, and Francisco Pereira. VICE: Variational interpretable concept embeddings. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[21] Daniel J. Navarro and Thomas L. Griffiths. Latent features in similarity judgments: A nonparametric bayesian approach. *Neural Computation*, 20(11):2597–2628, 2008. doi: 10.1162/neco.2008.04-07-504.

[22] Brett D. Roads and Michael C. Mozer. Obtaining psychological embeddings through joint kernel and metric learning. *Behavior Research Methods*, 51:2180—-2193, 2019. doi: 10.3758/s13428-019-01285-3.

[23] Brett D. Roads and Bradley C. Love. Enriching ImageNet with human similarity judgments and psychological embeddings. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3547–3557, 6 2021. doi: 10.1109/CVPR46437.2021.00355.

[24] Roger N Shepard. Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22(4):325–345, 1957.

[25] Roger N Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27(2):125–140, 1962.

[26] Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*, 2011.

[27] Catherine Wah, Grant Van Horn, Steve Branson, Subhransu Maji, Pietro Perona, and Serge Belongie. Similarity comparisons for interactive fine-grained categorization. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, 6 2014.

[28] Quentin F Gronau and Michael D Lee. Bayesian inference for multidimensional scaling representations with psychologically interpretable metrics. *Computational Brain & Behavior*, 3:322–340, 2020.

[29] Robert M. Nosofsky, Craig A. Sanders, Brian J. Meagher, and Bruce J. Douglas. Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*, 50:530–556, 2018. doi: 10.3758/s13428-017-0884-8.

[30] Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behavior*, pages 1173–1185, 2020. doi: 10.1038/s41562-020-00951-3.

[31] Robert Goldstone. An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, 26:381–386, 1994.

[32] Nikolaus Kriegeskorte and Marieke Mur. Inverse mds: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, 3:245, 2012. doi: 10.3389/fpsyg.2012.00245.

[33] Carl J Hodgetts, James O E Close, and Ulrike Hahn. Similarity and structured representation in human and nonhuman apes. *PsyArXiv*, Apr 2022. doi: 10.31234/osf.io/5vck6. URL psyarxiv.com/5vck6.

[34] Hirofumi Nakayama, Richard C. Gerkin, and Dmitry Rinberg. A behavioral paradigm for measuring perceptual distances in mice. *Cell Reports Methods*, 2(6):100233, 2022. ISSN 2667-2375. doi: https://doi.org/10.1016/j.crmeth.2022.100233. URL https://www.sciencedirect.com/science/article/pii/S2667237522001023.

[35] Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris Mcnorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37:547—-559, 2005. doi: https://doi.org/10.3758/BF03192726.

[36] J.P. Salmon, P.A. McMullen, and J.H. Filliter. Norms for two types of manipulability (graspability and functional usage), familiarity, and age of acquisition for 320 photographs of objects. *Behavior Research Methods*, 42(1):82–95, 2010. doi: 10.3758/BRM.42.1.82.

[37] Sebastian Bobadilla-Suarez, Christiane Ahlheim, Abhinav Mehrotra, Aristeidis Panos, and Bradley C Love. Measures of neural similarity. *Computational Brain & Behavior*, 3(4):369–383, 2020. doi: 10.1007/s42113-019-00068-5.

[38] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed

representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[39] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[41] Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 1 2015.

[42] Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020. doi: 10.1073/pnas.1907367117.

[43] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.

[44] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983. doi: https://doi.org/10.1016/S0364-0213(83)80009-3.

[45] Amos Tversky and J. Wesley Hutchinson. Nearest neighbor analysis of psychological

spaces. *Psychological Review*, 93(1):3–22, 1986. doi: https://doi.org/10.1037/0033-295X.93.1.3.

[46] Ronald R Coifman, Stephane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, Frederick Warner, and Steven W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences*, 102(21):7426–7431, 2005.

[47] Brian K. Lee, Emily J. Mayhew, Benjamin Sanchez-Lengeling, Jennifer N. Wei, Wesley W. Qian, Kelsie Little, Matthew Andres, Britney B. Nguyen, Theresa Moloy, Jane K. Parker, Richard C. Gerkin, Joel D. Mainland, and Alexander B. Wiltschko. A principal odor map unifies diverse tasks in human olfactory perception. *bioRxiv*, 2022. doi: https://doi.org/10.1101/2022.09.01.504602.

[48] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1), 6 2014.

[49] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6338–6347. Curran Associates, Inc., 2017.

[50] Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Transactions*

*on Pattern Analysis and Machine Intelligence*, 44(12):10023–10044, 2022. doi: 10.1109/TPAMI.2021.3136921.

[51] Yuansheng Zhou, Brian H. Smith, and Tatyana O. Sharpee. Hyperbolic geometry of the olfactory space. *Science Advances*, 4(8):eaaq1458, 2018. doi: 10.1126/sciadv.aaq1458. URL `https://www.science.org/doi/abs/10.1126/sciadv.aaq1458`.

[52] Chad Giusti, Eva Pastalkova, Carina Curto, and Vladimir Itskov. Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences*, 112(44):13455–13460, 2015. doi: https://doi.org/10.1073/pnas.1506407112.

[53] Huanqiu Zhang, P. Dylan Rich, Albert K. Lee, and Tatyana O. Sharpee. Hippocampal spatial representations exhibit a hyperbolic geometry that expands with experience. *Nature Neuroscience*, 26:131–139, 2023. doi: 10.1038/s41593-022-01212-4.

[54] Yoshua Bengio, Jean-françcois Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Roux, and Marie Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL `https://proceedings.neurips.cc/paper/2003/file/cf05968255451bdefe3c5bc64d550517-Pap`

[55] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006. doi: 10.1109/CVPR.2006.100.

[56] David Badre, Apoorva Bhandari, Haley Keglovits, and Atsushi Kikumoto. The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, 38:20–28, 2021. doi: https://doi.org/10.1016/j.cobeha.2020.07.002.

[57] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D. Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893, 2019. doi: 10.1126/science.aav7893.

[58] Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, MORTEZA IBRAHIMI, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 2795–2823. Curran Associates, Inc., 2023.

[59] Scott Sievert, Daniel Ross, Lalit Jain, Kevin Jamieson, Rob Nowak, and Robert Mankoff. Next: A system to easily connect crowdsourcing and adaptive data collection. In *Proceedings of the 16th Python in Science Conference*, pages 113–119, 2017.

[60] Tim Valentine, Michael B Lewis, and Peter J Hills. Face-space: A unifying concept in face recognition research. *Quarterly Journal of Experimental Psychology*, 69(10): 1996–2019, 2016.

[61] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[62] Byron M Yu, John P Cunningham, Gopal Santhanam, Stephen Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for

low-dimensional single-trial analysis of neural population activity. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/file/ad972f10e0800b49d76fed33a21f6698-Pap

[63] Liwei Kuang, Fei Hao, Laurence T. Yang, Man Lin, Changqing Luo, and Geyong Min. A tensor-based approach for big data representation and dimensionality reduction. *IEEE Transactions on Emerging Topics in Computing*, 2(3):280–291, 2014. doi: 10.1109/TETC.2014.2330516.

[64] Ali Noroozi and Mansoor Rezghi. A tensor-based framework for rs-fmri classification and functional connectivity construction. *Frontiers in Neuroinformatics*, 14, 2020. ISSN 1662-5196. doi: 10.3389/fninf.2020.581897. URL https://www.frontiersin.org/articles/10.3389/fninf.2020.581897.

[65] Jihun Ham, Daniel D. Lee, Sebastian Mika, and Bernhard Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 47, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015417. URL https://doi.org/10.1145/1015330.1015417.

[66] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500): 2323–2326, 2000. doi: 10.1126/science.290.5500.2323. URL https://www.science.org/doi/abs/10.1126/science.290.5500.2323.

[67] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. doi: 10.1162/089976603321780317.

[68] David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[69] Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing*, 26(1):313–338, 2004. doi: 10.1137/S1064827502419154. URL `https://doi.org/10.1137/S1064827502419154`.

[70] John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, 2014. doi: https://doi.org/10.1038/nn.3776.

[71] Evren Gokcen, Anna I Jasper, João D Semedo, Amin Zandvakili, Adam Kohn, Christian K Machens, and Byron M Yu. Disentangling the flow of signals between populations of neurons. *Nature Computational Science*, 2:512–525, 2022. doi: https://doi.org/10.1038/s43588-022-00282-5.

[72] Tze Hui Koh, William E Bishop, Takashi Kawashima, Brian B Jeon, Ranjani Srinivasan, Yu Mu, Ziqiang Wei, Sandra J Kuhlman, Misha B Ahrens, Steven M Chase, et al. Dimensionality reduction of calcium-imaged neuronal population activity. *Nature*

*Computational Science*, pages 71–85, 2023. doi: https://doi.org/10.1038/s43588-022-00390-2.

[73] Juan A. Gallego, Matthew G. Perich, Lee E. Miller, and Sara A. Solla. Neural manifolds for the control of movement. *Neuron*, 94(5):978–984, 2017. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2017.05.025. URL `https://www.sciencedirect.com/science/article/pii/S0896627317304634`.

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

[75] Robert M Nosofsky. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115:39–57, 1986. doi: 10.1037/0096-3445.115.1.39.

[76] John K Kruschke. Alcove: an exemplar-based connectionist model of category learning. *Psychological Review*, 99(1):22–44, 1992.

[77] Bradley C Love, Douglas L Medin, and Todd M Gureckis. Sustain: a network model of category learning. *Psychological Review*, 111(2):309–332, 2004.

[78] Kurt Braunlich and Bradley C Love. Bidirectional influences of information sampling and concept learning. *Psychological Review*, 2021. doi: 10.1037/rev0000287.

[79] Christiane Ahlheim and Bradley C. Love. Estimating the functional dimensionality of neural representations. *NeuroImage*, 179:51–62, 2018. doi: https://doi.org/10.1016/j.neuroimage.2018.06.015.

[80] Michael L Mack, Alison R Preston, and Bradley C Love. Ventromedial prefrontal cortex compression during concept learning. *Nature communications*, 11(46):1–11, 2020. doi: 10.1038/s41467-019-13930-8.

[81] Michael L. Mack, Bradley C. Love, and Alison R. Preston. Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, 113(46):13203–13208, 2016. doi: 10.1073/pnas.1614048113.

[82] Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353, 2013. doi: 10.1109/TKDE.2012.51.

[83] Wieland Brendel, Ranulfo Romo, and Christian K Machens. Demixed principal component analysis. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL `https://proceedings.neurips.cc/paper/2011/file/f4a331b7a22d1b237565d8813a34d8ac-Pap`

[84] Dmitry Kobak, Wieland Brendel, Christos Constantinidis, Claudia E Feierstein, Adam Kepecs, Zachary F Mainen, Xue-Lian Qi, Ranulfo Romo, Naoshige Uchida, and Christian K Machens. Demixed principal component analysis of neural population data. *eLife*, 5:e10989, apr 2016. ISSN 2050-084X. doi: 10.7554/eLife.10989. URL `https://doi.org/10.7554/eLife.10989`.

[85] Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw,

Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013.

[86] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571:361–365, 2019. doi: https://doi.org/10.1038/s41586-019-1346-5.

[87] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017. doi: https://doi.org/10.1038/s41598-017-11873-y.

[88] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/cfcce0621b49c983991ead4c3d4d3b6b-Pap`

[89] Silvia Bernardi, Marcus K. Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C. Daniel Salzman. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4):954–967.e21, 2020. doi: https://doi.org/10.1016/j.cell.2020.09.031.

[90] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

[91] Nicholas J. Sexton and Bradley C. Love. Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Sci-*

*ence Advances*, 8(28):eabm2219, 2022. doi: 10.1126/sciadv.abm2219. URL `https://www.science.org/doi/abs/10.1126/sciadv.abm2219`.

[92] Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[93] Bradley C Love. Model-based fmri analysis of memory. *Current Opinion in Behavioral Sciences*, 32:88–93, 2020. doi: https://doi.org/10.1016/j.cobeha.2020.02.012.

[94] Roger N Shepard and Susan Chipman. Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, 1(1):1 – 17, 1970. doi: https://doi.org/10.1016/0010-0285(70)90002-2.

[95] Aarre Laakso and Garrison Cottrell. Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology*, 13(1):47–76, 2000. doi: 10.1080/09515080050002726.

[96] Harold Hotelling, Harold. RELATIONS BETWEEN TWO SETS OF VARIATES*. *Biometrika*, 28(3-4):321–377, 12 1936. doi: 10.1093/biomet/28.3-4.321.

[97] Hao-Ting Wang, Jonathan Smallwood, Janaina Mourao-Miranda, Cedric Huchuan Xia, Theodore D. Satterthwaite, Danielle S. Bassett, and Danilo Bzdok. Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists. *NeuroImage*, 216:116745, 2020. doi: https://doi.org/10.1016/j.neuroimage.2020.116745.

[98] Xiaowei Zhuang, Zhengshi Yang, and Dietmar Cordes. A technical review of canonical correlation analysis for neuroscience applications. *Human Brain Mapping*, 41(13): 3807–3833, 2020. doi: https://doi.org/10.1002/hbm.25090.

[99] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2019.

[100] Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

[101] Nello Cristianini, Jaz Kandola, Andre Elisseeff, and John Shawe-Taylor. On kernel target alignment. In *Innovations in machine learning*, pages 205–256. Springer, 2006.

[102] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13 (1):795–828, 2012.

[103] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 6 2019.

[104] *The effect of task and training on intermediate representations in convolutional neural networks revealed with modified RV similarity analysis*, 2019.

[105] Philipp Kaniuth and Martin N. Hebart. Feature-reweighted representational similarity analysis: A method for improving the fit between com-

putational models, brains, and behavior. *NeuroImage*, 257:119294, 2022. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2022.119294. URL https://www.sciencedirect.com/science/article/pii/S105381192200413X.

[106] Katherine R. Storrs, Tim C. Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, 33(10):2044–2064, 09 2021. ISSN 0898-929X. doi: 10.1162/jocn_a_01755. URL https://doi.org/10.1162/jocn_a_01755.

[107] Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

[108] A. Klami, S. Virtanen, and S. Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14(1):965–1003, 2013.

[109] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16 (12):2639–2664, 12 2004. doi: 10.1162/0899766042321814.

[110] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. *Proceedings of the 30th International Conference on Machine Learning*, 28(3):1247–1255, 2013.

[111] Daniela M Witten and Robert J. Tibshirani. Extensions of sparse canonical correlation

analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009. doi: doi:10.2202/1544-6115.1470.

[112] Jörn Diedrichsen, Atsushi Yokoi, and Spencer A. Arbuckle. Pattern component modeling: A flexible approach for understanding the representational structure of brain activity patterns. *NeuroImage*, 180:119–133, 2018. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.08.051. URL `https://www.sciencedirect.com/science/article/pii/S1053811917306985`. New advances in encoding and decoding of brain signals.

[113] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L Yamins, and James J DiCarlo. Brain-like object recognition with high-performing shallow recurrent anns. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 12805–12816. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/7813d1590d28a7dd372ad54b5d29d033-Pap`

[114] Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413 – 423, 2020. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2020.07.040. URL `http://www.sciencedirect.com/science/article/pii/S089662732030605X`.

[115] Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt,

Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/0937fb5864ed06ffb59ae5f9b5ed67a9-Pap`

[116] Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13890–13902. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/9f6992966d4c363ea0162a056cb45fe5-Pap`

[117] Brett D Roads and Bradley C Love. Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, 2(1):76–82, 2020. doi: 10.1038/s42256-019-0132-2.

[118] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822. PMLR, 2014.

[119] Martina Valente, Giuseppe Pica, Giulio Bondanelli, Monica Moroni, Caroline A. Runyan, Ari S. Morcos, Christopher D. Harvey, and Stefano Panzeri. Correlations enhance the behavioral readout of neural population activity in association cortex. *Nature Neuroscience*, 24:975—986, 2021. doi: https://doi.org/10.1038/s41593-021-00845-1.

[120] R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried.

Invariant visual representation by single neurons in the human brain. *Nature*, 435: 1102–1107, 2005.

[121] Jacob S Prince and Talia Konkle. Neural and computational evidence that category-selective visual regions are facets of a unified object space. *Journal of Vision*, 22: 4428–4428, 2022. doi: https://doi.org/10.1167/jov.22.14.4428.

[122] Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, 110(7):1258–1270, 2022.

[123] Ramon Nogueira, Chris C. Rodgers, Randy M. Bruno, and Stefano Fusi. The geometry of cortical representations of touch in rodents. *Nature Neuroscience*, 2023. doi: 10.1038/s41593-022-01237-9.

[124] Olivia Guest and Bradley C Love. What the success of brain imaging implies about the neural code. *Elife*, 6:e21397, 2017. doi: 10.7554/eLife.21397.

[125] Apoorva Bhandari, Christopher Gagne, and David Badre. Just above chance: is it harder to decode information from prefrontal cortex hemodynamic activity patterns? *Journal of Cognitive Neuroscience*, 30(10):1473–1498, 2018.

[126] Jingfeng Zhou, Chunying Jia, Marlian Montesinos-Cartagena, Matthew P. H. Gardner, Wenhui Zong, and Geoffrey Schoenbaum. Evolving schema representations in orbitofrontal ensembles during learning. *Nature*, 590:606–611, 2021. doi: https://doi.org/10.1038/s41586-020-03061-2.

[127] Stefano Fusi, Earl K Miller, and Mattia Rigotti. Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37:66–74, 2016. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2016.01.010. URL `https://www.sciencedirect.com/science/article/pii/S0959438816000118`. Neurobiology of cognitive behavior.

[128] Tatyana O Sharpee. An argument for hyperbolic geometry in neural circuits. *Current Opinion in Neurobiology*, 58:101–104, 2019. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2019.07.008. URL `https://www.sciencedirect.com/science/article/pii/S0959438818302411`. Computational Neuroscience.

[129] Andrea Guidolin, Mathieu Desroches, Jonathan D. Victor, Keith P. Purpura, and Serafim Rodrigues. Geometry of spiking patterns in early visual cortex: a topological data analytic approach. *Journal of The Royal Society Interface*, 19(196):20220677, 2022. doi: 10.1098/rsif.2022.0677. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2022.0677`.

[130] Adam N Hornsby and Bradley C Love. Sequential consumer choice as multi-cued retrieval. *Science Advances*, 8(8):eabl9754, 2022.