Estimation of perceptual scales using ordinal embedding

Siavash Haghiri Department of Computer Science, University of Tübingen, Germany

Felix A. Wichmann Department of Computer Science, University of Tübingen, Germany

Ulrike von Luxburg Department of Computer Science, University of Tübingen, Germany
Max Planck Institute for Intelligent Systems, Tübingen, Germany

In this article, we address the problem of measuring and analyzing sensation, the subjective magnitude of one's experience. We do this in the context of the method of triads: The sensation of the stimulus is evaluated via relative judgments of the following form: "Is stimulus S_i more similar to stimulus S_j or to stimulus S_k?" We propose to use ordinal embedding methods from machine learning to estimate the scaling function from the relative judgments. We review two relevant and well-known methods in psychophysics that are partially applicable in our setting: nonmetric multidimensional scaling (NMDS) and the method of maximum likelihood difference scaling (MLDS). Considering various scaling functions, we perform an extensive set of simulations to demonstrate the performance of the ordinal embedding methods. We show that in contrast to existing approaches, our ordinal embedding approach allows, first, to obtain reasonable scaling functions from comparatively few relative judgments and, second, to estimate multidimensional perceptual scales. In addition to the simulations, we analyze data from two real psychophysics experiments using ordinal embedding methods. Our results show that in the one-dimensional perceptual scale, our ordinal embedding approach works as well as MLDS, while in higher dimensions, only our ordinal embedding methods can produce a desirable scaling function. To make our methods widely accessible, we provide an R-implementation and general rules of thumb on how to use ordinal embedding in the context of psychophysics.

Introduction

The quantitative study of human behavior dates back to at least 1860, when the experimental physicist Gustav Theodor Fechner published *Die Elemente der Psychophysik* (Fechner, 1860). Since Fechner's seminal work, the "measurement of sensation magnitude"—nowadays typically referred to as "psychophysical scaling"—has been one of the central aims of psychophysics (Gescheider, 1988).[1] Psychophysical scaling is formally defined as the problem of quantifying the magnitude of sensation induced by a physical stimulus (Marks & Gescheider, 2002; Krantz, Luce, Suppes, & Tversky, 1971).

In the following, we assume that there exists a physical quantity—the external stimulus—that we can objectively measure. The perception (or sensation, the subjective or internal experience) of the stimulus, however, is usually hard to measure and quantify. The (difference) scaling problem refers to experiments and methods designed to find the functional relation between the perceived (internal) magnitude and the (external) stimulus. An example of a scaling function is shown in Figure 1. In this figure, the physical stimulus S and its perceived counterpart ψ are denoted on the x- and y-axes, respectively. Throughout the rest of the article, we refer to this function as the scaling function.

Traditional scaling methods

Early attempts to obtain the scaling function by Fechner were based on the concatenation of just-noticeable-difference (JND), the smallest amount of change in the stimulus level that is noticeable by a human observer. Fechner assumed that each JND in S corresponds to one fixed-size unit of the perceptual scale ψ and attempted to reconstruct the

Citation: Haghiri, S., Wichmann, F. A., & von Luxburg, U. (2020). Estimation of perceptual scales using ordinal embedding. *Journal of Vision*, *20*(9):14, 1–20, https://doi.org/10.1167/jov.20.9.14.

https://doi.org/10.1167/jov.20.9.14 Received July 3, 2019; published September 21, 2020 ISSN 1534-7362

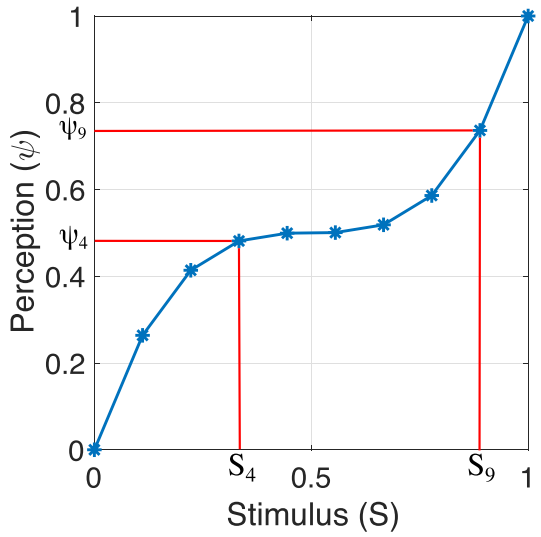

Figure 1. An example of a scaling function. The x-axis shows the physical stimulus values (S) with 10 discrete steps. The y-axis denotes the perceived value (ψ).

scaling function based on this assumption (Fechner, 1860; Luce & Edwards, 1958). Fechner thus tried to link *discriminability* and *subjective magnitude* in a simple way. However, the Fechnerian approach—albeit sometimes successful—has been vigorously criticized for both theoretical and empirical reasons and cannot serve as a generic method to obtain scaling function (e.g., Norris & Oliver, 1898; Stevens, 1957; Gescheider, 1988). Thurstonian scaling is an alternative approach proposed to solve the scaling problem in the tradition of linking discriminability to subjective magnitude, incorporating an internally variable mapping from stimulus to sensation (internal "noise" in modern parlance) (Thurstone, 1927). Thurstonian scaling is based on discrimination of stimuli pairs. The perceptual distance of two stimuli is determined by the probability that a human observer can discriminate them. However, like Fechner's JND approach, Thurstonian scaling is criticized because discriminability is, at best, only *indirectly* and in yet to be understood ways related to sensory magnitude (Krantz, 1972; Stevens, 1961).

Another well-known approach to scaling, but this time not based on discriminability, is termed *direct* magnitude estimation (Stevens, 1957). In this approach, a human observer is asked to provide intensity values corresponding to physical stimuli in a way that ratios of given values represent the ratios of perception. However, Shepard pointed out that there might exist an unknown and undesirable *response transformation function* that the direct magnitude estimation method neglects (Shepard, 1981).

For a much more detailed and in-depth overview and discussion of the traditional psychophysical scaling methods, we refer the reader to (Gescheider, 1988).

Scaling and the method of triads

An alternative approach to data acquisition—neither based on JND-style discrimination nor on direct magnitude estimation—is based on triplet comparisons (Torgerson, 1958). This approach is often referred to as *method of triads* in the psychophysics literature. Based on a fixed discretization of the physical stimulus, say $S_1, \ldots, S_n$, the method of triads asks participants to make comparisons of the following form: "Is stimulus S_i more similar to stimulus S_j or to stimulus S_k?" In the computer science and machine learning literature, such a question is called a **triplet question** (or, interchangeably, a triplet comparison).

Rather than attempting accurate quantitative measurements of a particular phenomenon, triplet questions aim at qualitative (ordinal) observations. The obvious potential of such an approach is that the statements do not depend as much on the response transformation function of the observers and that the issue of scaling answers across many observers becomes easier. In addition, studies in the machine learning literature indicate the robustness of the triplet comparison approach (Demiralp, Bernstein, & Heer, 2014; Li, Malave, Song, & Yu, 2016). The obvious challenge of the method of triads is how we can use the participants' answers to estimate the scaling function. More precisely, we need to estimate the magnitudes of perception $\hat{\psi}_1, \ldots \hat{\psi}_n$ in a way that is consistent with the answers to the queried triplet questions.

Let us give an example. Consider a psychophysical "slant-from-texture" experiment that has been designed to find the functional relation of the perceived angle to the true angle of a tilted flat plane with a dotted texture (Rosas, Wichmann, & Wagemans, 2004; Rosas, Ernst, Wagemans, & Wichmann, 2005; Rosas, Wichmann, & Wagemans, 2007; Aguilar, Wichmann, & Maertens, 2017). Figure 2 (top) shows the various stimuli used in the experiment by Rosas et al. (2004) and Aguilar et al. (2017). The bottom, left image of Figure 2 depicts an example of a triplet question designed for this task. The participant is asked, "Which of the two bottom images, S_j or S_k, is more similar to the top image S_i?" Based on the answers to a set of such triplet questions, the goal is then to reconstruct the scaling function that describes the relation of perceived angle ψ and the slant degree S. Figure 2 (bottom, right) shows the function that has been estimated with the t-distributed stochastic triplet embedding (t-STE) method described below.

The approach of triplet comparisons—the method of triads—is not new to psychophysics; there has

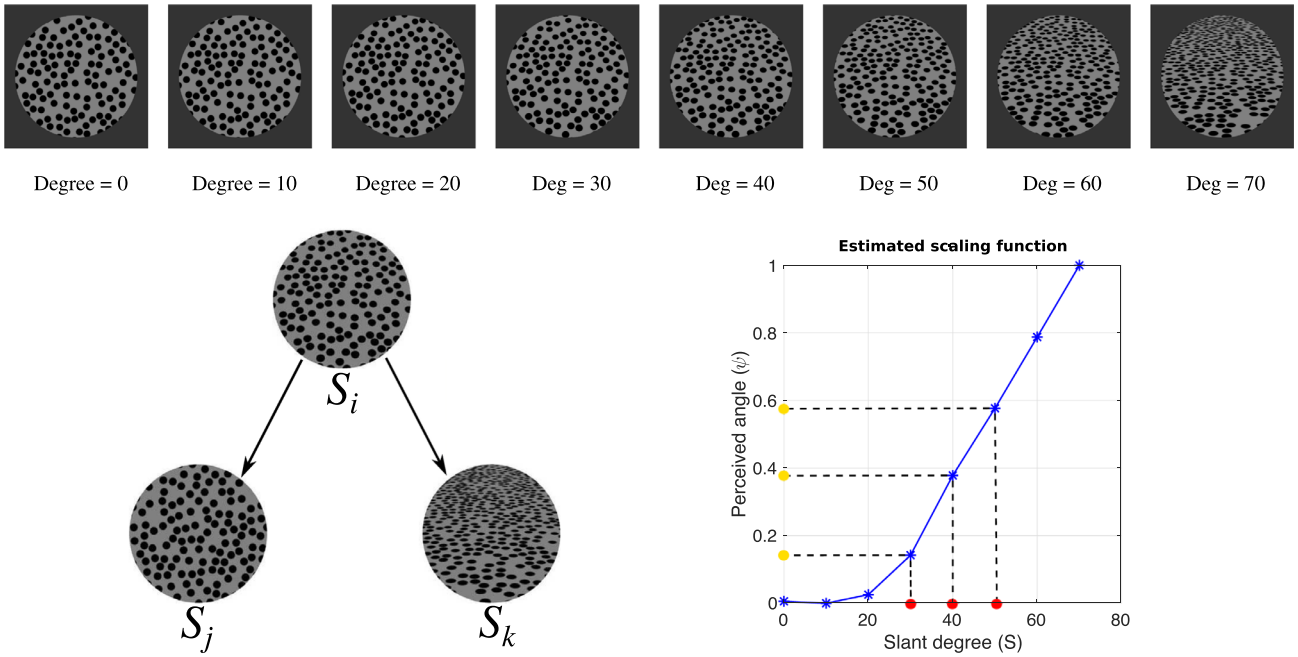

Figure 2. Top: Eight stimuli used in the slant-from-texture experiment (Aguilar et al., 2017). Bottom, left: An example of a triplet question used for the experiment. The triplet question is: "Which of the bottom images, S_j or S_k, is more similar to the top image S_i?" Bottom, right: The scaling function estimated by the comparison-based embedding method (*t*-STE). The red points on the x-axis correspond to three stimuli (S), while the yellow points on the y-axis represents their perceived values (ψ). In the second section (Embedding methods), we describe in detail, how the position of yellow points corresponds to the ordinal embedding from the triplet questions.

been a very long tradition in psychology to explore methods to estimate perceptual (difference) scales from clearly visible *supra-threshold* differences in stimulus appearance (Torgerson, 1958; Coombs, Dawes, & Tversky, 1970; Marks & Gescheider, 2002). There exists another line of research that estimates the perceptual function based on the comparisons between two pairs of stimuli levels (Schneider, 1980; Schneider, Parker, & Stein, 1974). The simple coordinate adjustment technique can estimate the one-dimensional representations of perceptual scales if it has access to the answers to all comparisons between stimuli pairs (Schneider, 1980).

Recently, a more generic approach, called maximum likelihood difference scaling (MLDS), has become popular in vision science (Maloney & Yang, 2003; Knoblauch & Maloney, 2010). There have been reports that both naive as well as seasoned observers find the method of triads with supra-threshold stimuli intuitive and fast, requiring less training (Aguilar et al., 2017; Wichmann et al., 2017) than for the more traditional methods in psychophysics such as direct magnitude estimation or, in particular, methods based on JNDs.

While clearly attractive, MLDS has one limitation, however. It can only be used to estimate one-dimensional scaling functions, that is, it cannot deal with cases when perception is intrinsically multidimensional (e.g., color perception). On the other hand, the evaluation of comparison-based data has been an active field of research in computer sciences and machine learning (Schultz & Joachims, 2003; Agarwal et al., 2007; Tamuz, Liu, Belongie, Shamir, & Kalai, 2011; Ailon, 2011; Jamieson & Nowak, 2011; Van Der Maaten & Weinberger, 2012; Kleindessner & von Luxburg, 2014; Terada & von Luxburg, 2014; Ukkonen, Derakhshan, & Heikinheimo, 2015; Arias-Castro, 2015; Jain, Jamieson, & Nowak, 2016; Haghiri, Ghoshdastidar, & von Luxburg, 2017). The core question of these studies is to use the answers to triplet comparisons to find a Euclidean representation of the items (in our case, psychophysical stimuli). This problem is systematically studied in the machine learning literature under the name of **ordinal embedding**. A number of fast and accurate algorithms have been developed to solve the ordinal embedding problem (Agarwal et al., 2007; Van Der Maaten & Weinberger, 2012; Terada & von Luxburg, 2014). As we will show in this article, these algorithms may also be useful in psychophysics, vision science and the cognitive sciences in general.

This article is organized as follows: In the second section (Embedding methods), we review two traditional psychophysical scaling methods, nonmetric multidimensional scaling (NMDS) and MLDS, that

are used to analyze data from triplet comparisons. We then introduce the ordinal embedding problem of the machine learning literature and discuss its advantages in comparison to the traditional embedding methods of psychophysics. The third section (Simulations) is dedicated to extensive simulations comparing the performance of ordinal embedding to the applicable competitors in psychophysics. In the fourth section (Experiments), we examine the ordinal embedding methods in two real psychophysics experiments. In the fifth section (How to apply ordinal embedding methods in psychophysics), we provide instructions and rules of thumb on how to use the comparison-based approach and the ordinal embedding algorithms in psychophysics experiments. In the last section, we conclude the article by discussing the advantages of the ordinal embedding for scaling problem and mentioning some of the open problems.

Embedding methods

NMDS

NMDS by Shepard and Kruskal is a well-established method to analyze dissimilarity data (Shepard, 1962; Kruskal, 1964a, 1964b). It assumes that a complete matrix of dissimilarities (not necessarily metric distances) between pairs of items is given. We denote the dissimilarity of items i and j by δ_{ij}. In the context of psychophysics, this matrix usually comes from a human (psychophysical) experiment. Shepard posed the problem of estimating a d-dimensional Euclidean representation of items, say $y_1, y_2, \dots y_n \in \mathbb{R}^d$, such that the pairwise distances of estimates are consistent with a monotonic transform of the given dissimilarities. Key to the method is that it only takes the rank order of the dissimilarities into consideration. This is attractive in many psychophysics experiments where the magnitude of dissimilarities cannot be quantitatively measured, whereas the rank order of distances is considered more reliable—the same argument that we have made earlier in favor of ordinal embedding (see above).

If $d_{ij} = \|y_i - y_j\|$ is the Euclidean distance of the embedded items y_i and y_j in $\mathbb{R}^d$, then the quality of a Euclidean representation is measured by a quantity called **stress** (Kruskal, 1964b):

$$\text{stress} = \frac{\sum_{ij} (d_{ij} - g(\delta_{ij}))^2}{\sum_{ij} {d_{ij}}^2}, \quad (1)$$

where g is a monotonic function to be determined. The smaller the stress, the better the Euclidean representation. The numerator measures the squared loss between the transformed input dissimilarities $g(\delta_{ij})$ and the Euclidean distances d_{ij}. By minimizing the stress, we try to achieve that the distances d_{ij} are as close as possible to the monotonic transform of dissimilarities $g(\delta_{ij})$. The role of the denominator is to prevent the degenerate solution where the d_{ij} all converge to 0.

The goal of NMDS is to find the Euclidean representation of items that minimizes the stress function, where g can be chosen from the set of all monotonic transform functions. The approach by Kruskal (1964a) finds an estimation of the optimal solution through an iterative two-step optimization procedure. In the first step, a configuration of embedding points $y_1, y_2, \dots y_n$ is fixed; this means that the distance values d_{ij} are also fixed. Then a greedy algorithm is suggested (later called isotonic regression) to find the monotonic function g that minimizes the stress function. In the second step of optimization, the values of $g(\delta_{ij})$ are fixed and the embedding points $y_1, y_2, \dots y_n$ are adjusted by a gradient descent algorithm to minimize the stress. The two steps are repeated iteratively until the stress value shows no further improvement or it becomes smaller than a certain threshold.

The NMDS algorithm has been used extensively in psychology (Reed, 1972; Smith & Ellsworth, 1985; Barsalou, 2014), neuroscience (de Beeck, Wagemans, & Vogels, 2001; Kayaert, Biederman, de Beeck, & Vogels, 2005; Kaneshiro, Guimaraes, Kim, Norcia, & Suppes, 2015), and broader fields (Liberti, Lavor, Maculan, & Mucherino, 2014; Machado, Mata, & Lopes, 2015). The nonparametric flavor of the method makes it a general-purpose algorithm that is easy to apply. In addition, it can find representations in multidimensional spaces. However, the algorithm has a major drawback: As described above, the algorithm needs the *full* dissimilarity matrix as input. Alternatively, in a setting of triplet comparisons, one can also implement the algorithm with just the knowledge on the ranking (ordering) of *all* the distance values δ_{ij}. This ordering can of course be computed from triplet questions, but it requires in order of $n^2 \log n$ triplet questions to sort all pairwise distances. This property makes NMDS infeasible for many applications in psychophysics, as the number of required triplet comparisons grows very quickly with the number of stimuli. For example, consider the case of $n = 15$ stimuli. There exist $m = \binom{15}{2} = 105$ dissimilarity values. NMDS requires the ordering of all these dissimilarity values. It is well known in computer science that to order m items, we need to ask about $m \log m$ comparisons; thus, $m \log m = 105 \cdot \log_2 105 \approx 700$ triplet questions are required to run NMDS. For ordinal embedding methods, the number of required triplet comparisons depends on the embedding dimension, and it has been proven to be of the order $dn \log_2 n$. If we embed in a two-dimensional space, then this amounts to $2 \cdot 15 \cdot \log_2 15 \approx 120$ triplet comparisons.

The difference becomes more drastic with larger n. If we assume $n = 50$, using the same calculation, NMDS requires about 12,570 triplet comparisons, whereas ordinal embedding methods require only about 570 triplet comparisons.

A second disadvantage of NMDS is that the optimization algorithm tries to solve a highly nonconvex optimization problem and typically gets stuck in a local but not the global minimum of the stress function. This local optimum can be arbitrarily far off from the global optimum. However, to be fair, this nonconvex optimization problem is shared with most of the other algorithms considered in this article.

Many variants of NMDS exist. As an example, consider (Ramsay, 1977; Takane, 1978) a situation in which the authors make a number of explicit model assumptions and then apply a maximum likelihood approach. This approach has been adapted for the analysis of data gathered by the method of triads (Bonnardel et al., 2016), where the triad responses are used to estimate the similarity matrix between items directly. The main differences between these more statistical approaches and machine learning might, up to some point, be of a philosophical nature: Rather than making many explicit model assumptions (for example in Ramsay (1977), a lognormal noise model, explicit weights on coordinates, powers of Euclidean distances to deal with non-Euclidean data), machine learning algorithms try to operate with minimalist assumptions — because they tend to be applied to data that rarely satisfy statistical model assumptions. This mind-set also makes it necessary to take care of overfitting: Rather than finding the best maximum of a likelihood function, machine learning takes into account aspects of robustness (as in t-STE) or, for example, the large margin principle, as in soft ordinal embedding (SOE). As another consequence, machine learning models typically operate with one big optimization problem rather than splitting the task into several separate steps (such as first estimating distances and, in a second step, constructing an embedding) the rationale being that each intermediate estimation step is yet another source of error and overfitting.

MLDS

Decades after the introduction of NMDS, MLDS was proposed to solve a specific instance of the difference scaling problem (Knoblauch, Charrier, Cherifi, Yang, & Maloney, 1998; Maloney & Yang, 2003; Krantz et al., 1971). Originally, MLDS asked quadruplet questions that involve four stimulus levels. If we denote the perceptual scale of four stimuli S_i, S_j, S_k, S_l by $\psi_i, \psi_j, \psi_k, \psi_l$, then a quadruplet question asks whether the difference in perception $|\psi_i - \psi_j|$ is larger or smaller than the difference of perception $|\psi_k - \psi_l|$. Note, however, that triplet questions are indeed a subset of quadruplet questions, implying that the MLDS method is also applicable to triplet questions.

The MLDS model (Maloney & Yang, 2003) assumes that the perceptual scale is a scalar value denoted by ψ—in the ordinal embedding language, it always embeds in a one-dimensional space. In contrast to NMDS, the MLDS method uses a parametric model. For a quadruplet of stimulus levels S_i, S_j, S_k, S_l, for simplicity denoted by $(i, j; k, l)$, a decision random variable is defined as

$$Dec(i, j; k, l) = |\psi_i - \psi_j| - |\psi_k - \psi_l| + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a zero-mean Gaussian noise with standard deviation $\sigma > 0$. If $Dec(i, j; k, l) > 0$, then the observer would respond that the pair (i, j) has a larger difference than the pair (k, l). In this case, the response to the quadruplet $q = (i, j; k, l)$ is set to $R_q = 1$; otherwise, the response is $R_q = 0$. The goal of the MLDS is now to estimate the perception scale ψ that maximizes the likelihood of the observed quadruplet answers. We first set $\psi_1 = 0, \psi_n = 1$ to remove degenerate solutions. Now, assuming that $R_1, R_2, \ldots, R_m \in \{0, 1\}$ denote the independent responses to m quadruplet questions, the likelihood of the perceptual scales given the quadruplet answers is

$$\begin{aligned}&\mathcal{L}(\psi_2, \ldots, \psi_{n-1}, \sigma | R_1, \ldots, R_m)\\ &= \prod_{q=1}^{m} \Phi(\Delta_q)^{R_q}[1 - \Phi(\Delta_q)]^{1-R_q},\end{aligned}$$

where $\Phi(.)$ denotes the cumulative distribution function of $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and $\Delta_q = |\psi_i - \psi_j| - |\psi_k - \psi_l|$ for the quadruplet $q = (i, j; k, l)$.[2] As it is the case for the stress function of NMDS, the likelihood function of MLDS is not convex with respect to the perceptual scale values ψ_i. Thus, the proposed numerical methods to maximize this likelihood might get stuck in a local maximum.

Knoblauch and Maloney (2010) introduced a more generalized version of MLDS, based on the generalized linear model (GLM), that also accepts triplet questions as inputs.

There are a number of advantages of the MLDS method: From a theoretical point of view, the maximum likelihood estimator is unbiased and has minimum variance among the unbiased estimators (however, these nice properties only hold for the global maximum of the likelihood function, not the local one, possibly discovered by the actual algorithm). It has been shown empirically that a reasonably small subset of quadruplets is enough to construct good scaling functions. Finally, it has been demonstrated that the

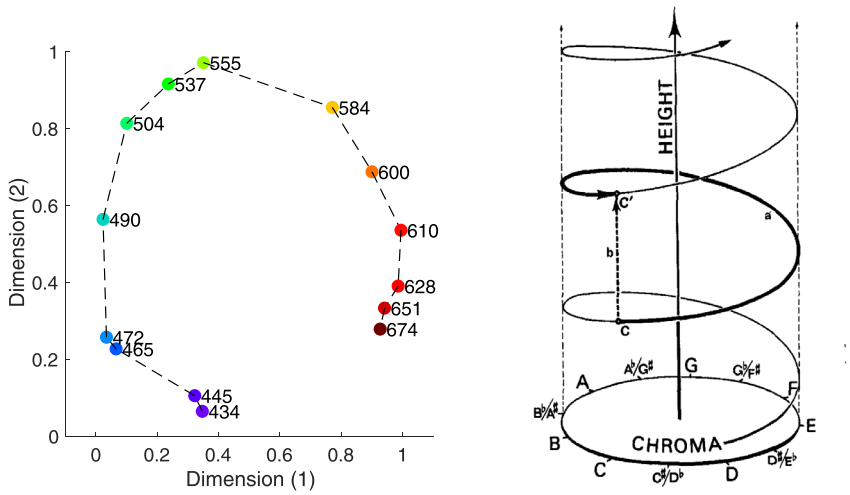

Figure 3. Left: The two-dimensional circle of color perception gathered by similarity measurements between 14 colors (Shepard, 1962). The *wavelength* of each color is written on the right side of the colored dot. Note that we have reconstructed the color circle with NMDS based on the original dissimilarity data. Right: The helix proposed by Shepard for the pitch perception. The physical stimulus, that is, pitch, varies along the spiral path of the curve and the three-dimensional space describes the perception (Shepard, 1982).

variance of the output behaves reasonably with respect to the input noise level (Maloney & Yang, 2003), and MLDS can be interpreted within the framework of signal detection theory (Green & Swets, 1966).

However, MLDS has one major drawback: The algorithm only works for one-dimensional perceptual spaces. In some cases (see the examples of color and pitch perception in Figure 3), the scales definitely need more than one dimension. To assess whether one-dimensional spaces might be suitable for a given data set at hand, Knoblauch and Maloney (2010) suggest a certain goodness-of-fit criterion (the "six-point condition"). If this condition is not satisfied, this might indicate the requirement of embedding in higher dimensions.

There have been attempts to extend the MLDS model to the case of multidimensional perceptual scales. For instance, Radonjić, Cottaris, and Brainard (2019) proposed a model to study the perception of objects in two dimensions, color and material, but they need to make many strong assumptions. As we explain in the next section, the ordinal embedding methods of machine learning can deal with the multidimensional perceptual case in general, without making assumptions such as independence of perceptual scales, additive noise, and independence of noise and stimulus.

Ordinal embedding

General setup

The comparison-based setting has recently become popular in machine learning literature (Schultz & Joachims, 2003; Agarwal et al., 2007; Van Der Maaten & Weinberger, 2012; Amid & Ukkonen, 2015; Ukkonen et al., 2015; Balcan, Vitercik, & White, 2016). Instead of stimulus levels, machine learning deals with a set of abstract items, say $x_1, x_2, \ldots, x_n$, that come from some abstract space $\mathcal{X}$. Furthermore, we assume that there exists a dissimilarity function $\delta : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ that describes the dissimilarity of the items. Often, in machine learning, we assume that δ is symmetric but not necessarily a metric. In our current setting in psychophysics, we assume that the function δ is not available, yet we have access to an oracle that responds to a triplet question $t = (i, j, k)$, based on the unknown dissimilarity. The triplet question will be "Is item x_i more similar to item x_j or item x_k"? We denote this triplet question by $t = (i, j, k)$, and sometimes we will call the point x_i the pivot of the triplet. The response to the triplet is denoted by R_t and stored as the following:

$$R_t = \begin{cases} 1 & \text{if the oracle responds that } x_j \text{ is more similar to } x_i \\ -1 & \text{if the oracle responds that } x_k \text{ is more similar to } x_i \end{cases} \quad (2)$$

Assume that the answers to a subset of triplet questions $T \subset \{(i, j, k) | x_i, x_j, x_k \in \mathcal{X}\}$ are collected from the oracle. Given an embedding dimension d and the answers to the triplet questions T, the ordinal embedding aims to find points $y_1, y_2, \ldots y_n \in \mathbb{R}^d$ in a d-dimensional Euclidean space such that the Euclidean distances are consistent with the answers of the queried triplet questions. The consistency of an embedding with

respect to triplet $t = (i, j, k)$ can be judged as follows:

$$R_t \cdot sgn(\|y_i - y_k\|^2 - \|y_i - y_j\|^2) = \begin{cases} 1, & \text{if the embedding is consistent with } R_t \\ -1, & \text{if the embedding is not consistent with } R_t \end{cases}$$

where function *sgn* returns the sign of a real value. The goal of **ordinal embedding** is to find an embedding $y_1, ..., y_n$ that maximizes the number of consistent triplets. Intuitively, we would like to solve the following optimization problem:

$$\max_{y_1,...,y_n \in \mathbb{R}^d} \sum_{t=(i,j,k)\in T} R_t \cdot sgn(\|y_i - y_k\|^2 - \|y_i - y_j\|^2). \quad (3)$$

However, this formulation leads to a number of algorithmic obstacles. From a mathematical point of view, it is not always possible to find a perfect d-dimensional embedding for an arbitrary dissimilarity function δ. Moreover, in a practical setting, the answers to the triplets might be noisy. Therefore, the optimal solution is not necessarily consistent with the full set of triplets T. And finally, as written above, the objective function is discrete-valued, which makes it even harder to optimize. For all these reasons, various adaptations of the stress function and optimization heuristics have been suggested to address these problems. Below we describe one particular algorithm in more detail.

Connection to the scaling problem

Ordinal embedding solves the scaling problem of psychophysics in the following way: The different stimuli S_i play the same role as the abstract items x_i in the ordinal embedding problem, and the perception values ψ_i correspond to the embeddings y_i. Concretely, given a standard scaling function as in Figure 2 (bottom right), the ordinal embedding output corresponds to the positions of the perception values on the y-axis (yellow points in Figure 2, bottom right). Thus, given the ordinal embedding output (y-values) and the values of the physical stimuli, we can reconstruct the scaling function.

Consider again the example of the slant-from-texture problem in Figure 2: Given the slant stimuli $S_1, ..., S_n$, participants were asked a number of triplet questions involving the stimuli $S_1, ..., S_n$. Then, we provided the answers of these triplet questions to an ordinal embedding algorithm and asked the algorithm to construct a one-dimensional embedding. This resulted in the yellow points $y_1, ...y_n$ on the y-axis (in the plot, we only marked three out of eight stimuli with yellow points to keep it simple). These points can now be identified as the perception values $\psi_1, ..., \psi_n$, so we can finally draw the scaling function by connecting the points (S_i, ψ_i). More details on this experiment are provided in the Experiments section.

While in the example of the slant experiment, we used a one-dimensional embedding, ordinal embedding methods can also construct a *multidimensional* embedding that describes the perceptual space of humans. Let us discuss two examples that demonstrate why this might be important. One famous example is *color perception.* Figure 3 (left) shows the two-dimensional color circle proposed by Shepard and Ekman (Shepard, 1962; Ekman, 1954). The figure has been constructed with the NMDS algorithm based on a 14×14 similarity judgment matrix. The wavelength of each color is written at the right side of each colored dot. In our context, the important observation is that human observers perceive the violet colors with short wavelengths as similar to the red colors with long wavelengths. This suggests a circular perceptual internal space, which can only be realized in at least two dimensions. A second example is *pitch perception* of sounds. Even though auditory frequency is again one-dimensional, the pitch is perceived along a three-dimensional helix (Shepard, 1982; Houtsma, 1995). Figure 3 (right) shows the proposed perception space by Shepard. In both cases, pitch and color, multidimensional ordinal embedding is the tool that can enable a researcher to find perceived values in higher-dimensional Euclidean spaces, which might be necessary to properly capture the similarity structure of perception (or cognition).

Stochastic triplet embedding

In recent years, there has been a surge of methods to address the ordinal embedding problem in the machine learning community, for example, generalized nonmetric multidimensional scaling (Agarwal et al., 2007), the crowd-median kernel (Tamuz et al., 2011), stochastic triplet embedding (STE) (Van Der Maaten & Weinberger, 2012), and SOE (Terada & von Luxburg, 2014). In general, the focus of the machine learning community is to build methods that require only a small number of triplets to embed a large number of items, make as few assumptions as possible, and be robust toward noise in the data.

In the following, we focus on one particular class of methods, STE, and its variant, t-STE, because in our experience, they work very well and are based on a simple model that is also plausible in a psychophysics setting. The STE method introduces the probabilistic model defined in Equation 4 to solve the ordinal embedding problem. Assume that $y_1, \ldots, y_n \in \mathbb{R}^d$ were the correct representations of our objects. The model assumes that if participants are being asked whether y_i is closer to y_j or to y_k, then they give a positive answer

Method	Data required	Statistical noise model	Multidimensional
NMDS	Complete order of all distances	No	Yes
MLDS	Partial set of quadruplets (or triplets)	Yes	No
t-STE	Partial set of triplets	Yes	Yes

Table 1. The comparison of ordinal embedding methods. Each row corresponds to one method, while the properties are listed in the columns.

with probability

$$p_{ijk} = \frac{\exp(-\|y_i - y_j\|^2)}{\exp(-\|y_i - y_j\|^2) + \exp(-\|y_i - y_k\|^2)}. \quad (4)$$

Intuitively, "easy" triplet questions (where the distances $\|y_i - y_j\|$ and $\|y_i - y_k\|$ are very different) will be answered correctly in most of the cases, whereas difficult triplet questions (where $\|y_i - y_j\|$ is about as large as $\|y_i - y_k\|$) can often be mixed up. Given the answers to a set of triplets, the STE algorithm attempts to maximize the likelihood of the embedding point configuration with respect to the answered triplets. If the answer to a triplet question $t = (i, j, k)$ is given according to Equation 2, and if we assume that triplet questions are answered independently, the likelihood of an embedding given the answers to a set of triplets T is given as

$$\mathcal{L}\left(y_1, \ldots, y_n | R_1, \ldots, R_{|T|}\right) = \prod_{t=(i,j,k)\in T, R_t=1} p_{ijk} \cdot \prod_{t=(i,j,k)\in T, R_t=-1} (1 - p_{ijk}).$$

The log-likelihood is maximized to find the solution of ordinal embedding. In the above formulation, the probability of satisfying a triplet goes rapidly to zero when the difficulty of a triplet question increases. As a result, severe and slight violations of a triplet are penalized almost the same. To make the statistic more robust, the authors propose to replace the Gaussian functions $\exp(-\|y_i - y_j\|^2)$ with Student-t functions with a heavier tail kernel (Van Der Maaten & Weinberger, 2012). The modified method is called t-distributed STE (t-STE).

This algorithm can deal with a large number of items (stimulus levels) and reasonable number of triplets, and it is robust to noise, which is an important characteristic when dealing with psychophysics data. Unlike MLDS, the algorithm is capable of embedding in higher-dimensional Euclidean spaces. However, as with all the other methods, the proposed optimization problem is not convex, which makes it vulnerable to inappropriate local optima.

Summary of embedding methods

In Table 1, we summarize the properties of the different embedding methods. The ordinal embedding methods can produce high-quality results with a small set of triplet answers. This property makes them superior to traditional NMDS that requires the full order of distances. On the other hand, the ordinal embedding methods are not limited to the case of one-dimensional functions, as it is the case for MLDS.

As the number of items (and consequently the number of triplets) grows, many of the ordinal embedding algorithms become computationally slow. This is, however, more of a concern for machine learning purposes, where we deal with thousands of items and hundreds of thousands of triplets. For standard psychophysics experiments, ordinal embedding algorithms such as STE and t-STE have an acceptable running time. Our experiments are performed on an iMac 18.3 (2017) with a 3.4-GHz i5 quad-core processor. On this machine, the (t)-STE algorithm, implemented in MATLAB, requires about 30 min to embed 100 items in two dimensions using 2,000 triplet answers. As this analysis needs to run only once after all the triplets of all participants have been recorded, we do not think that this is a problem in a typical psychophysical setting.

Simulations

In this section, we describe simulations that compare ordinal embedding algorithms with the corresponding approaches in psychophysics (NMDS and MLDS).

Simulation setup

Stimulus and perceptual scale

We assume that the stimulus lives on a scale from 0 to 1, and the true relation between the physical stimulus and the perception is encoded by a function $f : (0, 1) \to (0, 1)^d$, where the dimension d of the perceptual space will typically be 1 or 2. We consider n uniformly chosen steps for the stimulus levels, denoted by $S = \{S_1, S_2, \ldots, S_n\}$. In our simulations, we assume

that a true perceptual scale exists, for the stimulus S_i is denoted by $y_i = f(S_i)$. We will choose different functions f for our different simulations below.

Generating subsets of triplet questions

In order to have a fair comparison, we will provide the same number of triplet questions to each of our algorithms. There are some subtle differences in the implementations that we need to pay attention to. MLDS (as in Knoblauch & Maloney, 2010) assumes that the ordering between input stimuli is known. Given three input stimuli S_i, S_j, and S_k, it looks up their relative order (e.g., $S_i < S_j < S_k$) and then always asks the triplet question $\psi_j - \psi_i > \psi_k - \psi_j$ (with ψ_j as the "pivot point" and without absolute values). Ordinal embedding methods, on the other hand, do not assume any knowledge on the stimulus ordering (and there might be cases where the latter does not exist, for example, if we consider multimodal inputs). Instead, given three input stimuli S_i, S_j, and S_k, it can consider three different triplet questions: $|\psi_j - \psi_i| > |\psi_k - \psi_i|$ (here ψ_i is the pivot point) and the corresponding questions with the other two as the anchor points.

Consequently, n stimuli levels give rise to $\binom{n}{3}$ valid triplet questions for MLDS and $3\binom{n}{3}$ valid triplet questions for the ordinal embedding methods. In all our simulations, we provide the same number of triplets to all embedding algorithms. A random subset of triplets is chosen uniformly without replacement from the set of valid triplets for each algorithm, where this set of valid triplets is slightly different for MLDS and the other algorithms, as described above. In our simulations, the size of the random subset of triplets will be chosen in the range $r \cdot \binom{n}{3}$ with $r \in \{0.1, 0.2, 0.4, 1\}$. The value $r = 1$ is equivalent to choosing the whole set of valid triplets for the MLDS method and a third of the set of valid triplets for the other methods.

Underlying model to generate triplet answers for MLDS and ordinal embedding

In order to simulate answers to the triplet questions, we construct a model that resembles a typical observer of a psychophysical experiment. Given a fixed perceptual scale function f, we assume that the simulated observer answers the triplet questions based on a noisy version of this function, denoted by $\tilde{y}_i = f(S_i) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma \cdot I_d)$ is a zero-mean Gaussian noise with unit covariance matrix and standard deviation σ in d dimensions. In our simulations, we use σ in the range of $\{0.01, 0.05, 0.1, 0.5\}$. The process of generating triplets for the ordinal embedding methods is as follows. The simulated observer produces the answer to the queried triplet question $t = (i, j, k)$ (S_i being the pivot of the triplet comparison) by

$$R_t = \begin{cases} 1 & \text{if } \|\tilde{y}_i - \tilde{y}_j\| < \|\tilde{y}_i - \tilde{y}_k\|, \\ -1 & \text{otherwise.} \end{cases}$$

Note again that the embedding values y play the same role as the perceptual scale values ψ in the psychophysics notation. We sometimes use a different notation to emphasize that the embedding values y can be multidimensional and to make a clear distinction to scalar values of ψ. The scalar ψ is depicted in Figure 1 and also in the MLDS method.

As mentioned in the previous subsection, the GLM formulation of MLDS requires a slightly different way for asking triplet questions. The triplet answer for three stimuli levels $S_i < S_j < S_j$ is constructed according to the following equation:

$$R_t = \begin{cases} 1 & \text{if } \tilde{y}_j - \tilde{y}_i < \tilde{y}_k - \tilde{y}_j, \\ -1 & \text{otherwise.} \end{cases}$$

Providing triplet answers to the algorithms. The above mentioned model produces answers to the triplet questions, for the methods based on triplet answers, namely, SOE, (t)-STE, and MLDS.

Underlying model to create the input to NMDS

NMDS requires dissimilarities between all pairs of items but in the end only makes use of the order between these values due to the monotonic transformation function g used in the definition of stress in Equation 1. For our simulations, we generate a set of noisy perceptual values for n stimuli levels as before, $\tilde{y}_i = f(S_i) + \epsilon$, and then explicitly compute all dissimilarity values $\delta_{i,j} = \|\tilde{y}_i - \tilde{y}_j\|$. These are then the values that we give to the NMDS algorithm. Note that because NMDS requires the full matrix of dissimilarities, we only apply and compare the NMDS algorithm to the other methods when $r = 1$, that is, all triplet questions are being asked. This procedure makes sure that all three algorithms get the same amount of information.

Embedding methods

We now apply various algorithms to generate embeddings or perceptual scales. For STE and t-STE, we use the MATLAB implementation by Van Der Maaten and Weinberger (2012).[3] We use the default optimization parameters for both methods. The degree of freedom for the t-Student kernel is set to $\alpha = 1$ for the t-STE method. We also use the R-implementation of a second algorithm from the machine learning community, SOE, with the default parameter settings.

For MLDS, we use the R-package available on the CRAN repository,[5] again with the default optimization parameter settings. For the NMDS algorithm, we use the MATLAB implementation, which is available by calling the function "mdscale." The implementation optimizes the stress function defined by Kruskal (1964a); see Equation 1.

In all cases, we set the embedding dimension to the dimension of true perceptual function. In the section on real experiments, we also consider cases where the embedding dimension is not known.

All embedding methods solve a nonconvex optimization problem and thus are prone to find inaccurate local optima. To reduce this effect, we run all the algorithms 10 times with random initializations. Among the 10 embedding outputs, we choose the one that has the smallest triplet error (see next subsection for a definition).

Independent of the above repetition, which is supposed to reduce the effect of local minima, each embedding method is executed 10 times, on 10 independent draws of the random input data. This repetition is meant to analyze the statistical behavior of the algorithm, average, and the variance. We plot the average values over these 10 repetitions in the main article and provide the standard deviations in the supplementary material (so the figures are not overly cluttered).

Evaluating the results

We consider two approaches to evaluate the performance of the various methods:

(1) Mean squared error (MSE): For one-dimensional perceptual spaces where the ground truth is known, we can compute the MSE between the estimated scales $\hat{y}$ and the true perceptual function values y. However, we need to be careful as the embedding results are only unique up to similarity transformations (scaling, rotation, and translation). So before computing the MSE, we need two steps of **normalization**. First, we transform the output of embedding to be in the range of (0,1) as our scaling functions are defined in this range (more precisely, we shift the minimum value to zero and divide all the values by the maximum). This takes care of translation and scaling. Second, the output is only unique up to rotation, which, in our one-dimensional scenario, consists of flipping the function values $\hat{y}$ to $-\hat{y}$ (note that if $\hat{y}$ satisfies all triplet questions, so does $-\hat{y}$). Therefore, we choose the one among $\hat{y}$ and $-\hat{y}$ that results in the smaller value of MSE. In this way, we choose the best rotation of the output.

(2) Triplet error: The MSE criterion is cumbersome to compute in multivariate scenarios, because we have to take into account all possible rotations of the embeddings. Moreover, in real-world scenarios, the MSE cannot be computed at all because the required, underlying ground truth is unknown. As an alternative, we propose to evaluate the quality of an embedding by its ability to predict the answers to (potentially new) triplet questions. To this end, we compute a quantity called the **triplet error**. Intuitively, the triplet error counts how many of the triplets are not consistently represented by the given embedding. Given an embedding $\hat{y}_1, ..., \hat{y}_n$ and a validation set T' of triplets, the triplet error of the embedding with respect to T' is defined as

$$\text{triplet error} = \frac{1}{|T'|} \sum_{t=(i,j,k)\in T'} \mathbb{1}\left\{R_t \cdot sgn(\|\hat{y}_i - \hat{y}_k\|^2 - \|\hat{y}_i - \hat{y}_j\|^2) = 1\right\}, \quad (5)$$

where the characteristic function $\mathbb{1}$ takes the value 1 if the expression in the curly parenthesis is true (that is, if the estimated embedding is not consistent with the new triplet t), and it takes the value 0 otherwise. Typically, the given set of answered triplets needs to be used both for constructing the embedding and for evaluating its quality. There are two ways to do this. The first, naive way is to set $T' = T$, meaning that we use the same set of triplets to construct the embedding and to measure its quality. In a second way, we perform k-fold cross-validation to avoid overfitting: We partition the set of input triplets T into k nonintersecting subsets ("folds"). We perform the embedding and the evaluation k times. In each iteration, we pick one of the folds as the validation set (T') and the rest of the folds as the training set (the input to the embedding algorithm). The final triplet error is the average over the triplet errors of the k validation sets. Throughout the rest of the article, we refer to the latter approach as **cross-validated triplet error**, while the first approach is simply called the **triplet error**.

One-dimensional perceptual space

Simulations with monotonic scales

Our first simulation involves a typical monotonic function as it occurs in many psychophysics experiments. The true perceptual function f (a sigmoid function) is shown in Figure 4a. Figures 4b, c shows the output embedding of the MLDS and STE algorithms for 10 iterations, respectively. The other ordinal embedding methods have a similar performance, and the output embeddings are reported in the supplementary material. The average (over 10 runs) MSE and triplet errors of various embedding algorithms are depicted in Figures 4d, e, respectively.

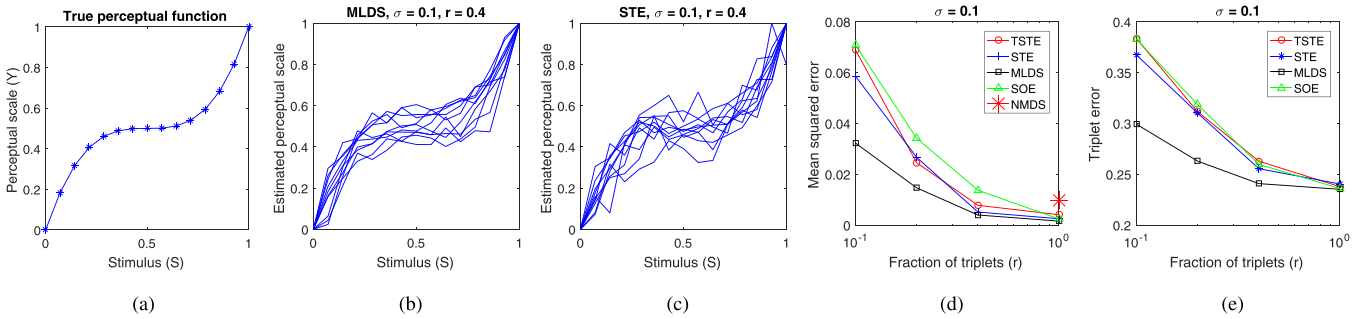

Figure 4. Comparison of various ordinal embedding methods (SOE, STE, *t*-STE) against the traditional embedding methods in psychophysics (MLDS and NMDS) for a monotonic one-dimensional perceptual function (Sigmoid). (a) The true perceptual function (y). (b) Ten embedding results ($\hat{y}$) of the MLDS method for a fixed value of standard deviation σ and triplet fraction r. (c) Ten embedding results ($\hat{y}$) of the STE method for a fixed value of standard deviation σ and triplet fraction r. (d) The average MSE of embedding methods. (e) The average triplet error of embedding methods.

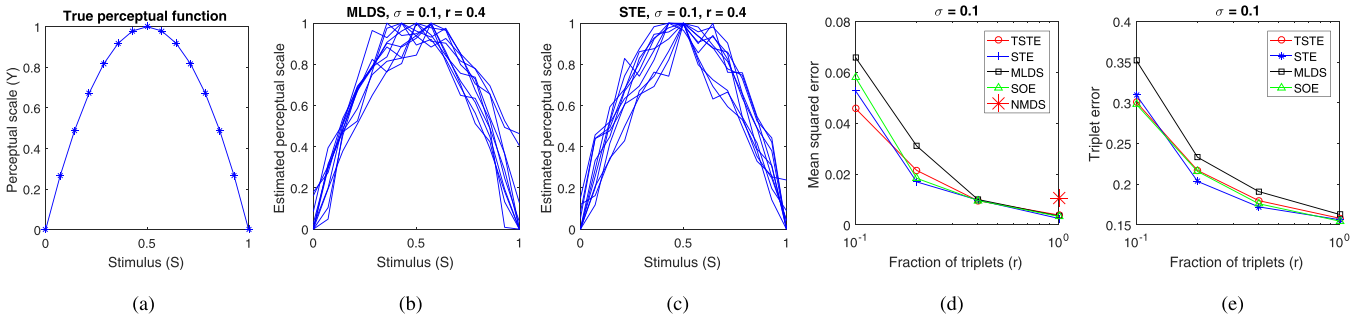

Figure 5. Comparison of various ordinal embedding methods (SOE, STE, *t*-STE) against the traditional embedding methods in psychophysics (MLDS and NMDS), for a nonmonotonic one-dimensional perceptual function (second-degree polynomial). (a) The true perceptual function (y). (b) Ten embedding results ($\hat{y}$) of the MLDS method for a fixed value of the standard deviation σ and triplet fraction r. (c) Ten embedding results ($\hat{y}$) of the STE method for a fixed value of standard deviation σ and triplet fraction r. (d) The average MSE of embedding methods. (e) The average triplet error of embedding methods.

In both error measures, the MLDS method performs slightly better than the ordinal embedding algorithms. More detailed results regarding this simulation, including the four ordinal embedding outputs and the performance of algorithms with other values of σ, can be found in the supplementary material; (see Figure 10). We also examine another monotonic function in Figure 11 of the supplementary material, and the results are consistent with the ones presented here.

Simulations with nonmonotonic scales

We now perform the same experiment with a nonmonotonic function: a second-degree polynomial function is chosen as the true perceptual function f; see Figure 5a. Figures 5b, c shows the output embedding of the MLDS and STE algorithms for 10 iterations, respectively (the embeddings produced by SOE and *t*-STE are quite similar to the STE; see supplementary material). The average (over 10 runs) MSE and triplet error of various embedding algorithms are depicted in Figures 5d, e, respectively.

The function shapes depicted in Figure 5b show the performance of the MLDS method for the nonmonotonic function. Considering the embeddings and two error plots, we can conclude that MLDS and ordinal embedding methods have very similar performance. MLDS performs slightly worse than ordinal embedding methods, when provided with fewer triplet answers (Figures 5d, e). In both monotonic and nonmonotonic cases, we observe that all methods converge to an output embedding with the same MSE and triplet error, when they are provided with a sufficiently large number of triplets.

Similar to the monotonic functions, we report the full details of the simulation in supplementary material; see Figure 13. We also perform the simulation on a sinosoid function. The results are quite similar to the second-degree polynomial function and are demonstrated in the Figure 12 of the supplementary material.

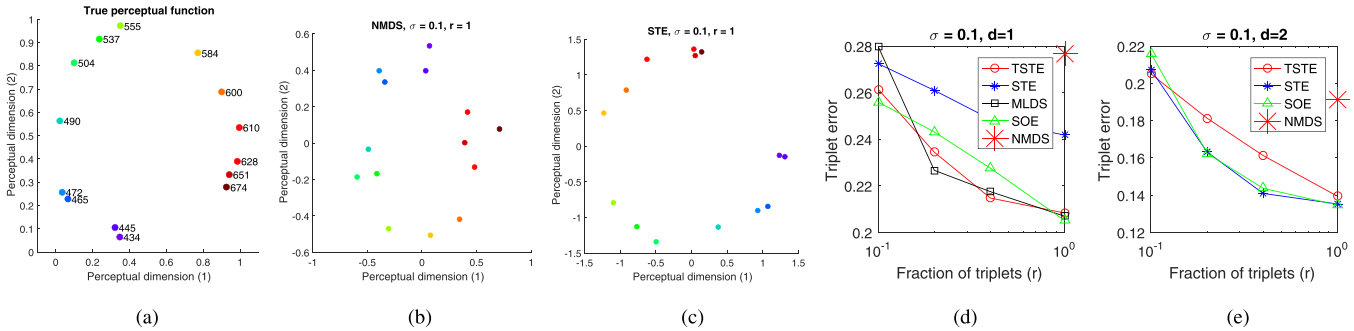

Figure 6. Comparison of the ordinal embedding methods (MLDS, STE, and *t*-STE) against the traditional NMDS method of psychophysics for the two-dimensional color perception function. (a) The true perceptual function in two dimensions. The stimulus value, color wavelength, is written beside each color. The two-dimensional vector space represents the perceptual space. (b) The embedding result of the NMDS method depicted in two dimensions for a fixed value of standard deviation σ and triplet fraction r. (c) The embedding result of the STE method depicted in two dimensions for a fixed value of standard deviation σ and triplet fraction r. (d, e) The average triplet error of various methods with embedding dimension $d = 1$ and $d = 2$, respectively.

Multidimensional perceptual space

So far, we considered simulations in which the perception could be represented in a one-dimensional Euclidean space. However, in some cases, such as the examples of color and pitch perception in Figure 3, more than one dimension is required to represent the perception. Here, we perform a simulation with a function mapping from one-dimensional stimulus space into a two-dimensional perceptual space.

In order to construct a realistic psychometric function f, we use the color similarity data[6] presented in Ekman (1954). We first construct a two-dimensional embedding using NMDS; see Figure 6a. In the following, this embedding will be considered our ground truth, which will then be used to generate further data (let us stress: we do not argue that this embedding is "correct" in any way; we just use it as a ground truth to generate further simulations).

To generate noisy triplets from our ground truth, we essentially proceed as before: We rescale the stimulus sizes (wavelengths) to the range of $S \in (0, 1)$ to be consistent with the underlying model we defined earlier. We define the true ground truth function f by our ground truth embedding, which is the true two-dimensional representation $y_i = f(S_i)$ of a stimulus S_i given by the values in Figure 6. Now we generate noisy versions $\tilde{y}_i$ of the perceptual scale functions, random subsets of triplets and noisy answers to triplet questions as described in the beginning of this section, and use the various algorithms to compute an estimated embedding $\hat{y}_i$ of all the stimuli. Note that the actual perceptual values are two-dimensional; therefore, the Euclidean distances of stimuli (color) values in the X-Y plane are used to produce the triplets.

Given the noisy triplets, we can now apply the diverse algorithms to the data. We generate embeddings in one dimension using all algorithms (MLDS, NMDS, SOE, STE, and *t*-STE), and embeddings in two dimensions using all algorithms except MLDS (which is not designed for this purpose).

We first study the difference between one- and two-dimensional embeddings. Figure 6d shows average triplet error of various embedding methods, when the embedding dimension is fixed to $d = 1$. In this setting, ordinal embedding algorithms and MLDS have very similar performance, all having a rather high triplet error (≈ 0.3 to ≈ 0.22). The comparison with triplet error of embedding in two dimensions (Figure 6e) reveals the difference. Indeed, the ability to embed in two dimensions enables the algorithm to improve the triplet error from ≈ 0.22 to ≈ 0.14, when $r = 1$. Thus, it is plausible that the proper dimension to represent the percept is (at least) two.

Next, we compare the results of the two-dimensional embedding algorithms. Figures 6b, c shows the two-dimensional embedding output of the NMDS and STE algorithms, respectively. The embeddings are shown for the parameter values $\sigma = 0.1$ and $r = 1$. The comparison of Figures 6b, c reveals the different performances of NMDS and ordinal embedding methods in the presence of noise. NMDS is known to be quite vulnerable to noise, and this can be seen from the figures as well. While STE produces a circle of colors fairly similar to the true perceptual function, the colors are somewhat mixed up in the NMDS embedding. The triplet error also shows that ordinal embedding algorithms outperform the NMDS method by a large margin—even though we have only half or less of the triplets available. More details regarding this experiment can be found in the supplementary material; see Figure 14.

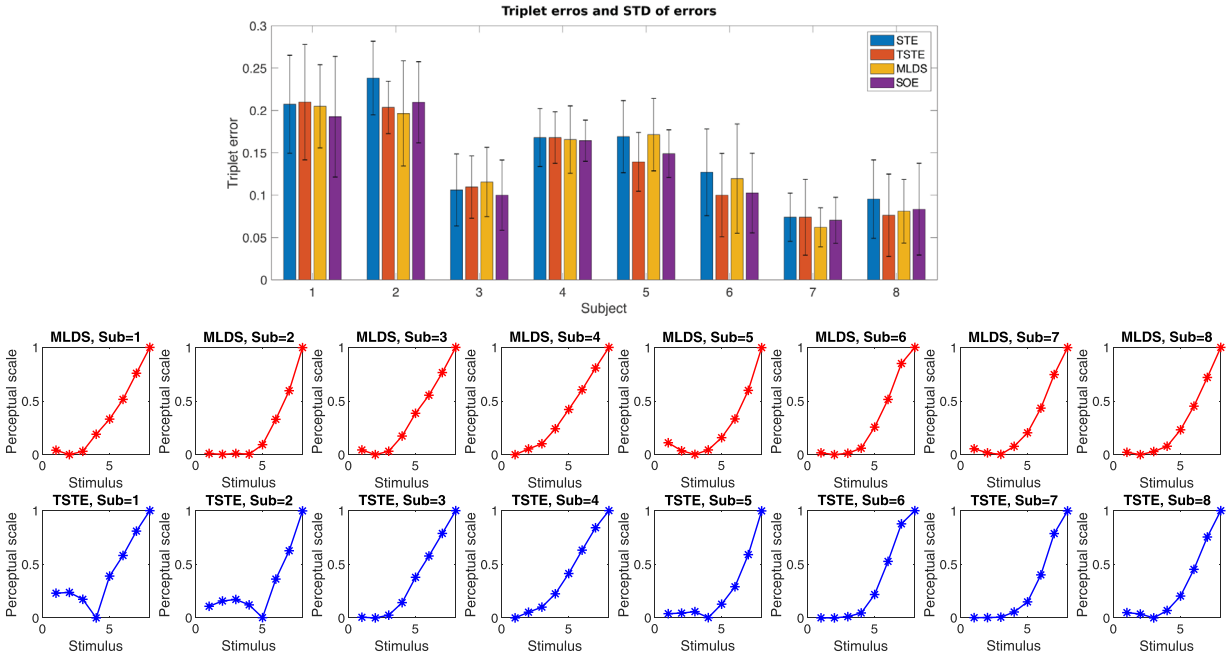

Figure 7. (Top) Average and standard deviation of cross-validated triplet error for eight subjects of the slant-from-texture experiment. Each group of bars shows the error for one subject, as each bar in the group corresponds to one of the embedding methods shown with different colors. (Bottom) The embedding outputs for eight subjects with two embedding methods: MLDS and *t*-STE. The MLDS method is depicted at the top row while the *t*-STE is shown at the bottom.

Experiments

In this section, we apply the comparison-based approach and ordinal embedding methods to two real experiments in visual perception: the slant-from-texture experiment that we have already mentioned above and a more complex "Eidolon" experiment.

Slant-from-texture experiment

This experiment intends to find the functional relation between the perceived angle of the slant with a dotted surface and the actual physical degree of slant. The data set was collected by Aguilar et al. (2017). Figure 2 (top) shows the eight stimuli used in this experiment. The degree of slant is varied from 0 to 70 degrees in steps of 10 degrees, making eight stimulus levels. Then participants had to answer triplet comparisons. As the experiment has initially been performed with the assumption of a monotonic relation of slant degree and the perception, for each combination of three stimuli $S_i < S_j < S_k$ (three degrees of tilting), only one triplet question has been asked: "Which of the two slant pairs have more difference (S_j, S_i) or (S_k, S_j)?" With eight levels of the stimulus, this results in $\binom{8}{3} = 56$ possible triplet questions. On this small order of magnitude, it is still possible to ask the participants to answer all possible triplet questions. Eight subjects participated in the study. Each subject has answered all 56 triplet questions several times, in order to reduce the effect of noisy responses. Subjects {1, 6, 8} have answered 420 triplet question in total, while the other subjects answered 840.

Since the ground truth embedding is unknown, we can only rely on the triplet error for evaluation of the embeddings. To avoid overfitting, we use 10-fold cross-validation to compute the *cross-validated triplet error* (see the definition in the simulation setup). Figure 7 (top) shows the average and standard deviation of the cross-validated triplet error for eight subjects and the four embedding methods: MLDS, STE, *t*-STE, and SOE. All algorithms have similar performance in this task.

In addition to the triplet error, we also show the embedding outputs of MLDS and *t*-STE for each of the eight subjects individually in Figure 7 (bottom). Note that these plots are generated with the full set of triplets, not only the training folds that are used to evaluate the triplet error. The resulting functions are similar, both across the two methods and across the participants. For some of the participants, we observe a noticeable difference between the embeddings of MLDS and *t*-STE, particularly Subjects 1 and 2. For these subjects, (*t*)-STE constructs a nonmonotonic function, while the MLDS function tends to be monotonic. The main reason for that is the nature of triplet questions. The

participants were asked, "Which of the two slant pairs are more different (S_j, S_i) or (S_k, S_j)?" If, for instance, the true perceptual function is decreasing from S_i to S_j, then the sign of $S_j - S_i$ is negative and a negative value should be evaluated with the GLM model (see Equation 9 in MLDS: Maximum Likelihood Difference Scaling in R (2010)). However, we believe that participants evaluated the differences based on the absolute value of the difference, that is, $|\psi_j - \psi_i|$. This can make a significant difference in the output embedding of MLDS. Although we performed the experiments considering the exact differences, in the real experiment, people have presumably answered the triplets with the notion of the absolute difference, as they were asked to indicate the pair with "more difference."

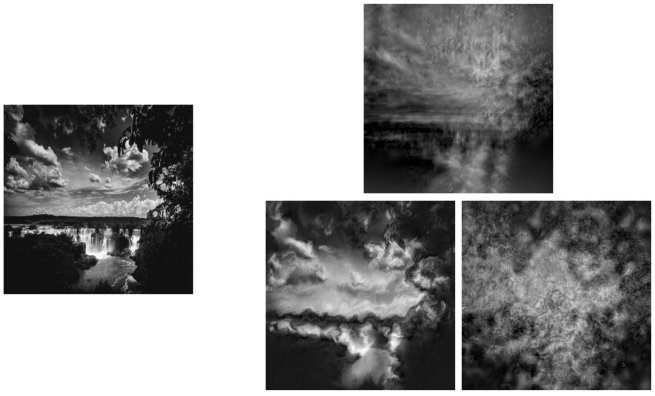

Figure 8. (Left) The original image in our Eidolon experiment. (Right) An example triplet question —: *"Which of the bottom two images is more similar to the top image?"*

The Eidolon experiment

Our final setup concerns a more "global" and less well-defined comparison of images. To generate "distorted" variants of images, we use the Eidolon Factory by Koenderink, Valsecchi, van Doorn, Wagemans, and Gegenfurtner (2017)—more specifically, its partially_coherent_disarray() function. In this toolbox, a given basis image can be distorted systematically using three different parameters called *reach*, *grain*, and *coherence*. An eidolon of a basis image corresponds to a parametrically altered version of this image. Reach controls the strength of a distortion (the higher the value, the stronger the amplification), grain modifies the fine-grainedness of the distortion (low values correspond to "highly fine-grained"), and a parameter value close to 1.0 for coherence indicates that "local image structure [is retained] even when the global image structure is destroyed" (Koenderink et al., 2017). From a perceptual point of view, we might want to know which and to what degree the image modifications influence the percept. Starting with a black and white image of a natural landscape as the basis image (see Figure 8, left), we generate 100 altered images, using reach and grain in {5, 12, 26, 61, 128} and coherence in {0.0, 0.33, 0.67, 1.0}. All possible combinations of these parameter values result in $5 \cdot 5 \cdot 4 = 100$ different images.

Lab experiment setup: In our lab, we asked three participants aged 19 to 25 to answer triplet questions; see Figure 8 (right) for an example question. For this purpose, participants use a standard computer mouse to click on one of the two bottom images that they deemed more similar to the top image. Stimuli were presented on a $1,920 \times 1,200$- pixel (484 × 302 mm) VIEWPixx LCD monitor (VPixx Technologies, Saint-Bruno, Canada) at a refresh rate of 120 Hz in an otherwise dark room. Viewing distance was 100 cm, corresponding to 3.66×3.66 degrees of visual angle for a single 256×256- pixel image. The surround of the screen was set to a gray value of 0.32 in the [0, 1] range, the mean value of all experimental images. The experiment was programmed in MATLAB (Release 2016a; The MathWorks, Inc., Natick, MA) using the Psychophysics Toolbox extensions version 3.0.12 (Brainard, 1997; Kleiner, Brainard, Pelli, Ingling, Murray, & Broussard, 2007) along with the iShow library of the Wichmann lab (http://dx.doi.org/10.5281/zenodo.34217).

Answers had to be given within 4.5 s after a triplet presentation onset; otherwise, the triplet was registered as unanswered and the experiment proceeded to the next triplet (this occurred in only 0.013% of all cases and can thus be safely ignored). Once a participant had answered a question, the next one appeared directly after a short fixation time of 0.3 s, during which only a white 20×20- pixel fixation rectangle at the center of the screen was shown. Before the experiment started, all test subjects were given instructions by a lab assistant and performed 100 practice trials to gain familiarity with the task. The set of practice triplets is disjoint from the set of experimental triplets. Participants were free to take a break every 200 triplet questions. They gave their written consent prior to the experiment and were either compensated €10 per hour for their time or gained course credit toward their degree. All test subjects were students and reported normal or corrected-to-normal vision.

Experiment design: Note that in our setup, there exist $n = 100$ stimuli (the different altered images), giving rise to around $3 \cdot \binom{n}{3} \approx 10^6$ possible triplet questions. In contrast to the previous slant-from-texture experiment with only eight stimuli, it is now absolutely impossible to ask a participant to evaluate all possible triplet questions. This already rules out the NMDS algorithm. This is now where the machine learning literature comes to aid: It has been proven theoretically that if the embedding dimension is d, then of the order $dn \log n$ triplet questions are sufficient to reconstruct the Euclidean representation of n items up to small

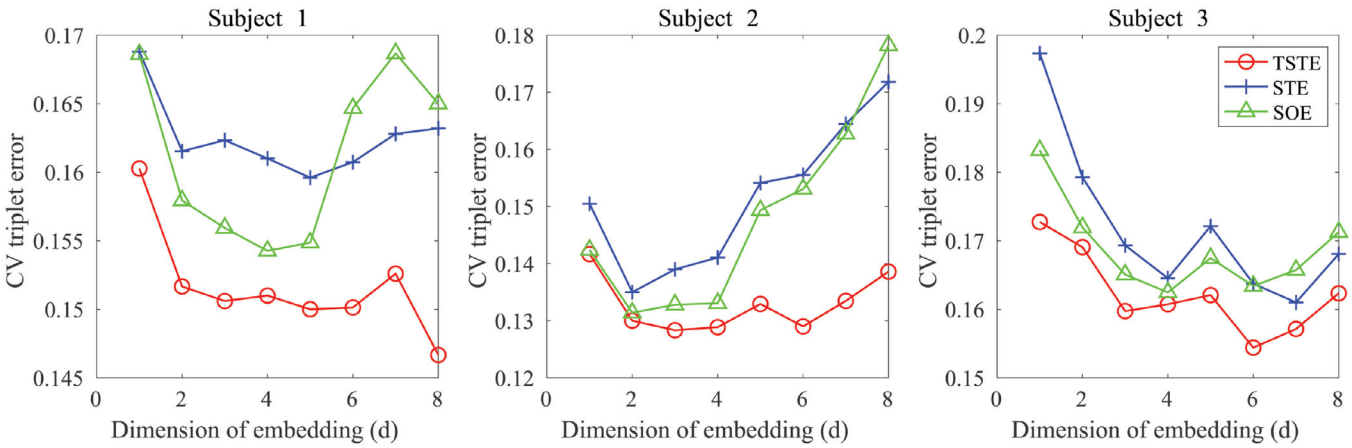

Figure 9. Cross-validated triplet error of three embedding methods for three subjects of the Eidolon experiment. Each plot corresponds to one subject, and each curve denotes the cross-validated triplet error of one method. The x-axis is the dimension of embedding.

error (Jain et al., 2016). Even though this is just an asymptotic statement and constants are completely ignored, it can serve as a guideline for our experimental setup. For example, assuming that the perceptual embedding dimension is not more than $d \approx 3$ (because three parameters are involved to modify the images), then $dn \log n = 100 \cdot 3 \cdot \log_2(100) \approx 2{,}000$ triplet questions should be enough to construct an accurate embedding. To be on the safe side, we hence decided to present 6,000 triplet questions to each participant of the experiment. These triplets have been chosen uniformly at random from the set of all possible triplets and were presented in three sessions of 2,000 triplets each.

Based on the triplet answers, we now run the ordinal embedding algorithms (STE, t-STE, SOE). As the best embedding dimension is unknown, we test dimensions in the range $d \in \{1, 2, \ldots, 8\}$. The MLDS method is performed only in the case of one dimension, as it is not applicable in multidimensional cases. We perform 10-fold cross-validation, and the cross-validated triplet error (see Equation 5) is reported as the evaluation criterion.

Figure 9 shows the cross-validated triplet error for three subjects with various dimensions and three embedding methods. Each plot corresponds to one subject, while each curve shows the error corresponding to one of the embedding methods. We can see that t-STE consistently outperforms the other methods. Note that the results of MLDS in case $d = 1$ are omitted from the plots: Even in dimension $d = 1$, the cross-validated triplet error for MLDS is larger than 0.25 for all three subjects, way larger than the error of the other methods. Thus, MLDS is not comparable to the performance of the best embedding methods and omitted from the plots—it would be off the scale in each of the panels. For all three subjects, increasing the embedding dimension from one to two definitely improves the embedding error—hence, we obviously need more than one dimension to describe the perceptual space. Adding further dimensions in most cases does not really help except perhaps for Subject 3. It looks as if further investigations and in particular more participants and a joint analysis over all participants would be necessary to come to a conclusion here if one wanted to know how the parameters of the Eidolon Factory are connected to perception.

The Eidolon experiment also points to another important methodological issue: The best embedding method (t-STE) leads to a cross-validated triplet error around 0.15—but is an error of 0.15 acceptable for this task? Could a (significantly) lower error be achieved if one, for example, collected more triplets? To answer this question, we would need to know the error baseline of human participants: There might be a proportion of ambiguous triplets, for example, for which no obviously "correct" answer exists. If, for example, we knew that 80% of the triplet questions had an easy, obviously correct answer, and 20% of the questions were so ambiguous that the answer was essentially random, then the best error rate we could hope for would be around 10%: On 80% of the triplets, we do not make any error, and on 20% of the triplets, we guess randomly, getting about 10% right and 10% wrong. Of course, in case of the Eidolon experiment, we do not have any external knowledge about the "difficulty" or "ambiguity" of triplets. But we can try to estimate it, and to this end, we conducted the following side experiment. We chose a set of 2,000 random triplets and asked each of them three times to each participant (triplets have been shuffled such that participants did not realize that they are answering the same triplets repeatedly). We now estimate the "difficulty" of a triplet by how consistent the repeated answers were: If a subject answers the same triplet question with different answers, we consider the question as "hard" and otherwise as "easy." We performed this experiment

with our three participants and they show the following percentage of hard triplets: 9.2%, 9.8%, and 11%. Consequently, a natural baseline for our embedding algorithms would be a triplet error of about 0.10. The cross-validated triplet errors reported in our plots above are actually close to this value, suggesting that our ordinal embeddings are close to what is achievable.

How to apply ordinal embedding methods in psychophysics

In order to make ordinal embedding methods more applicable for researchers who are unfamiliar with these methods, we now provide some basic rules of thumb.

How many triplets?

For a set of n stimuli, there exist $3\binom{n}{3}$ many triplet questions—already for moderate n, these are by far too many to ask to a participant of an experiment. However, for the ordinal embedding methods, a small subset of triplets already contains enough information to accurately reconstruct the true embedding. It has been proven that if the required embedding dimension is d, then of the order $\mathcal{O}(dn\log(n))$ triplets are sufficient to reconstruct the true embedding of n items (stimulus levels) up to a small error (Jain et al., 2016). According to this result, we suggest to start with a subset of size $dn\log(n)$ or $2dn\log(n)$ triplets and perform the ordinal embedding. If the time budget allows, one can still increase the number of triplets and see whether the error improves significantly, but $dn\log n$ should be a good baseline.

How to choose the subset of triplets?

Consider a set of n stimuli. At the first step, one needs to consider the whole set of possible triplets. As we mentioned earlier in simulations, every combination of three items from the stimuli set gives rise to **three** questions. Therefore, the complete set of possible triplet questions contains $3\binom{n}{3}$ triplets. The set of all possible triplets might be very large indeed, and thus a small subset of triplets needs to be subsampled. A natural question is: Which of the triplet questions among the whole set of possible questions should be chosen? Over the course of many years, we have tried many subsampling strategies in our group (Luxburg-lab): based on landmarks, based on active learning, based on estimated confidence values, based on the difficulty of triplet questions, and so on. However, in all our experiments, the simple strategy of selecting triplets uniformly at random from the set of all possible triplets performs surprisingly well in terms of triplet error. Hence, this is the strategy that we suggest to use.

How to evaluate the quality of the embedding?

We reported the MSE in our simulations; however, the true perceptual scale is not available in a real experiment. The general approach that we suggest for the evaluation of ordinal embedding is through the **cross-validated triplet error** (see Equation 5)—indeed, we suggest that this may be a good idea for MLDS and NMDS, too. The chosen subset of triplets needs to be partitioned into training and validation sets. The embedding method finds a Euclidean embedding for the perceptual scales, given the training set of triplets as input. We then calculate the cross-validated triplet error on the validation set. This procedure is preferable to the triplet error that is evaluated on the very same set that is used to construct the embedding; the latter can be highly biased and typically underestimates the true triplet error (overfitting).

How to choose the embedding dimension?

Note that from a formal point of view, increasing the embedding dimension can always lead to a decreasing triplet error—in the extreme case, it is always possible to embed n items in a space of $d = n$ dimensions without any error. But often it can be observed that there is a sharp decrease of the error as long as the dimension is still too small; once a sufficient dimension has been reached, the error decrease fades out (see the Eidolon example). We suggest to run the embedding algorithms in various dimensions, say from 1 to 10, and to choose the smallest dimension that shows an acceptable cross-validated triplet error. Also note that in some cases, it might also be possible to estimate the dimension of the data based on particular distance comparisons (Kleindessner & von Luxburg, 2015).

Which algorithm, which implementation?

Considering the results of the various algorithms on the many tasks and our experience in running ordinal embedding algorithms for many years, we consider t-STE as our method of choice. The original implementation of the authors is available at https://lvdmaaten.github.io/ste/Stochastic_Triplet_Embedding.html, implemented in MATLAB.

Discussion

In this article, we introduced ordinal embedding methods as a powerful approach to analyze triplet

comparisons gathered using the method of triads. Contrary to widespread belief, such methods provably require only a reasonably small ($dn \log n$) subset of triplet comparisons to achieve acceptable results for embedding n items. This property makes them preferable to traditional NMDS, which needs the rank order of all n^2 pairwise distances. Ordinal embedding methods are capable of embedding in multidimensional Euclidean spaces without restrictions on the scaling function. Thus, they have an advantage over MLDS, which is limited to one dimension. Furthermore, even in one-dimensional scaling scenarios, their performance is at least comparable to the one of MLDS. Hence, ordinal embedding methods such as t-STE are promising candidates to become a "default" psychophysical scaling algorithm.

Open issues

There are a few open issues regarding the use of ordinal embedding methods that need to be addressed in the future.

Confidence intervals

There have been considerable efforts to propose algorithms for the ordinal embedding problem. However, no particular study provides confidence intervals for the estimated embeddings. Although this issue is not taken very seriously in machine learning, for psychophysics, this is an issue of high importance. Some first steps in this direction have been taken in Lohaus, Hennig, and von Luxburg (2019), but there is definitely room for improvement.

Interpreting the embedding

A challenging yet important step is to interpret the embedding results. To make the point clear, consider the Eidolon experiment discussed in the previous section. After gathering a two-dimensional perceptual space and a mapping of stimuli in this space, there are a couple of natural questions arising. What does each perceptual dimension mean? How are the perceptual dimensions related to the parameters of the stimulus (in this case reach, coherence, and grain)? These are essential questions that can lead to better understanding of human perception.

Conjoint measurement

In addition to the general scaling problem, ordinal embedding is a promising candidate to tackle conjoint measurement problems (Luce & Tukey, 1964; Ho, Landy, & Maloney, 2008; Knoblauch & Maloney, 2012). In a conjoint measurement experiment, the sensory stimulus consists of more than one modality. Again we could ask participants to compare triplets of items and subsequently apply ordinal embedding. The approach of using triplet comparisons and ordinal embedding would need much less restriction than many of the approaches in conjoint measurement, which often rely on independence or additivity assumptions on the modalities.

Keywords: psychophysical scaling, maximum-likelihood difference scaling (MLDS), nonmetric multidimensional scaling (NMDS), psychophysics, ordinal embedding

Acknowledgments

The authors thank Robert Geirhos and Patricia Rubisch for the programming and running of the Eidolon experiment, Uli Wannek for the help with experimental setup, Guillermo Aguilar for providing the slant-from-texture data set and fruitful discussions, and Silke Gramer for administrative support. In addition, we thank Guillermo Aguilar and David-Elias Künstle for their feedback on our manuscript and our anonymous reviewers for their helpful reviews.

Supported by the German Research Foundation DFG via SFB 936/Z3, the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, DFG, ZUK 63), and the DFG Cluster of Excellence Machine Learning New Perspectives for Science, EXC 2064/1, project number 390727645. We also thank support of the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

Commercial relationships: none.
Corresponding author: Siavash Haghiri.
Email: siyavash.haghiri@gmail.com.
Address: Department of Computer Science, Neural Information Processing Group, Universität Tübingen, Tübingen, Germany.

Footnotes

[1] In fact, Plateau's work on relating physical stimulus magnitude to sensation precedes Fechner's but is not well known (Laming & Laming, 1996): We thank one of our reviewers for pointing us to Plateau's work. Other central aims are to measure detection and discrimination thresholds, or just-noticeable-differences (JNDs), reaction times (RT), and confidence ratings; see, for example, Wichmann and Jäkel (2018).
[2] In order to eliminate the absolute values, Maloney and Yang (2003) reorder the stimuli such that $\psi_i - \psi_j > 0$ and $\psi_k - \psi_l > 0$.
[3] https://lvdmaaten.github.io/ste/Stochastic_Triplet_Embedding.html.
[4] https://cran.r-project.org/web/packages/loe.
[5] https://cran.r-project.org/package=MLDS.
[6] https://faculty.sites.uci.edu/mdlee/similarity-data/.

References

Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., & Belongie, S. (2007). Generalized nonmetric multidimensional scaling. In M. Meila, & X. Shen (Eds.), *International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 11–18). San Juan, Puerto Rico: PMLR.

Aguilar, G., Wichmann, F. A., & Maertens, M. (2017). Comparing sensitivity estimates from MLDS and forced-choice methods in a slant-from-texture experiment. *Journal of Vision, 17*(1), 37, doi:10.1167/17.1.37.

Ailon, N. (2011). Active learning ranking from pairwise preferences with almost optimal query complexity. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems (NIPS)* (pp. 810–818).

Amid, E., & Ukkonen, A. (2015). Multiview triplet embedding: Learning attributes in multiple maps. In F. Bach, & D. Blei (Eds.), *International Conference on Machine Learning (ICML)* (pp. 1472–1480).

Arias-Castro, E. (2015). Some theory for ordinal embedding. *Bernoulli*, *23*(3), 1663–1693. Available from https://doi.org/10.3150/15-BEJ792, doi:10.3150/15-BEJ792.

Balcan, M., Vitercik, E., & White, C. (2016). Learning combinatorial functions from pairwise comparisons. In V. Feldman, A. Rakhlin, & O. Shamir (Eds.), *Conference on Learning Theory (COLT)* (*Vol. 49*, pp. 310–335). Columbia University, New York, New York, USA: PMLR.

Barsalou, L. W. (2014). *Cognitive psychology: An overview for cognitive scientists*. New York: Psychology Press.

Bonnardel, V., Beniwal, S., Dubey, N., Pande, M., Knoblauch, K., & Bimler, D. (2016). Perceptual color spacing derived from maximum likelihood multidimensional scaling. *Journal of the Optical Society of America A, 33*(3), A30–A36.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical Psychology*. Upper Saddle River, New Jersey: Prentice-Hall.

de Beeck, H. O., Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience, 4*, 1244.

Demiralp, Ç., Bernstein, M. S., & Heer, J. (2014). Learning perceptual kernels for visualization design. *IEEE Transactions on Visualization and Computer Graphics, 20*, 1933–1942.

Ekman, G. (1954). Dimensions of color vision. *Journal of Psychology, 38*, 467–474.

Fechner, G. T. (1860). *Elemente der psychophysik (elements of psychophysics)*. Leipzig: Breitkopf und Hrtel.

Gescheider, G. A. (1988). Psychophysical scaling. *Annual Review of Psychology, 39*, 169–200.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics (Vol. 1)*. New York, NY: Wiley.

Haghiri, S., Ghoshdastidar, D., & Luxburg, U. von. (2017). Comparison-based nearest neighbor search. In A. Singh, & J. Zhu (Eds.), *International Conference on Artificial Intelligence and Statistics (AISTATS)* (pp. 851–859). Fort Lauderdale, FL, USA: PMLR.

Ho, Y. X., Landy, M. S., & Maloney, L. T. (2008). Conjoint measurement of gloss and surface texture: Research article. *Psychological Science*, *19*(2), 196–204, doi:10.1111/j.1467-9280.2008.02067.x.

Houtsma, A. J. M. (1995). Pitch perception. *Hearing, 6*, 262.

Jain, L., Jamieson, K. G., & Nowak, R. (2016). Finite sample prediction and recovery bounds for ordinal embedding. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems (NIPS)* (pp. 2711–2719). Barcelona, Spain: Curran Associates, Inc.

Jamieson, K. G., & Nowak, R. D. (2011). Low-dimensional embedding using adaptively selected ordinal data. In *Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (p. 1077–1084). Monticello, IL, USA: IEEE.

Kaneshiro, B., Guimaraes, M. P., Kim, H.-S., Norcia, A. M., & Suppes, P. (2015). A representational similarity analysis of the dynamics of object processing using single-trial EEG classification. *PLoS One, 10*, 1–27.

Kayaert, G., Biederman, I., de Beeck, H. P. Op, & Vogels, R. (2005). Tuning for shape dimensions in macaque inferior temporal cortex. *European Journal of Neuroscience, 22*, 212–224.

Kleindessner, M., & von Luxburg, U. (2014). Uniqueness of ordinal embedding. In M. F. Balcan, V. Feldman, & C. Szepesvri (Eds.), *Conference on Learning Theory (COLT)* (pp. 40–67). Barcelona, Spain: PMLR.

Kleindessner, M., & von Luxburg, U. (2015). Dimensionality estimation without distances. In G. Lebanon, & S. V. N. Vishwanathan (Eds.),

International Conference on Artificial Intelligence and Statistics (AISTATS) (pp. 471–479). San Diego, California, USA: PMLR.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). Whats new in Psychtoolbox-3. *Perception, 36*, 1.

Knoblauch, K., Charrier, C., Cherifi, H., Yang, J., & Maloney, L. (1998). Difference scaling of image quality in compression-degraded images. Perception ECVP abstract, 27.

Knoblauch, K., & Maloney, L. T. (2010). MLDS: Maximum likelihood difference scaling in R. *Journal of Statistical Software, 25*, 1–26.

Knoblauch, K., & Maloney, L. T. (2012). Maximum Likelihood Conjoint Measurement. *Modeling Psychophysical Data in R. Use R!*, *32*, 229–256.

Koenderink, J., Valsecchi, M., Doorn, A. van, Wagemans, J., & Gegenfurtner, K. (2017). Eidolons: Novel stimuli for vision research. *Journal of Vision*, *17*(2), 7, doi:10.1167/17.2.7.

Krantz, D. (1972). Visual scaling. In J. D., & H. L.M. (Eds.), *Visual psychophysics* (pp. 660–689). Berlin, Heidelberg: Springer.

Krantz, D., Luce, D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement: Vol. 1. Additive and polynomial representations*. New York, NY: Academic Press.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*, 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29*, 115–129.

Laming, J., & Laming, D. (1996). J. Plateau: On the measurement of physical sensations and on the law which links the intensity of these sensations to the intensity of the source. *Psychological Research, 59*, 134–144.

Li, L., Malave, V. L., Song, A., & Yu, A. (2016). Extracting human face similarity judgments: Pairs or triplets? In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *CogSci* (pp. 1427–1432). Philadelphia, PA, USA.

Liberti, L., Lavor, C., Maculan, N., & Mucherino, A. (2014). Euclidean distance geometry and applications. *Siam Review, 56*, 3–69.

Lohaus, M., Hennig, P., & von Luxburg, U. (2019). Uncertainty estimates for ordinal embeddings. *Preprint available at Arxiv, abs/1906.11655*.

Luce, R. D., & Edwards, W. (1958). The derivation of subjective scales from just noticeable differences. *Psychological Review, 65*, 222.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1–27.

Machado, J., Mata, M., & Lopes, A. (2015). Fractional state space analysis of economic systems. *Entropy, 17*, 5402–5421.

Maloney, L. T., & Yang, J. N. (2003). Maximum likelihood difference scaling. *Journal of Vision*, *3*(8), 5, doi:10.1167/3.8.5.

Marks, L. E., & Gescheider, G. A. (2002). Psychophysical scaling. In J. Wixted (Ed.), *Stevens' handbook of experimental psychology (Vols. IV, Methodology in Experimental Psychology*. (p. 91–138). New York: John Wiley and Sons.

Norris, W. F., & Oliver, C. A. (1898). *System of diseases of the eye* (*Vol. 3*). Philadelphia: JB Lippincott.

Radonjíc, A., Cottaris, N. P., & Brainard, D. H. (2019). The relative contribution of color and material in object selection. *PLoS Computational Biology, 15*, 1–27.

Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika, 42*, 241–266.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382–407.

Rosas, P., Ernst, M. O., Wagemans, J., & Wichmann, F. A. (2005). Texture and haptic cues in slant discrimination: Reliability-based cue weighting without statistically optimal cue combination. *Journal of the Optical Society of America A, 22*, 801–809.

Rosas, P., Wichmann, F. A., & Wagemans, J. (2004). Some observations on the effects of slant and texture type on slant-from-texture. *Vision Research, 44*, 1511–1535.

Rosas, P., Wichmann, F. A., & Wagemans, J. (2007). Texture and object motion in slant discrimination: Failure of reliability-based weighting of cues may be evidence for strong fusion. *Journal of Vision*, *7*(6:3), 1–21, doi:10.1167/7.6.3.

Schneider, B. (1980a). Individual loudness functions determined from direct comparisons of loudness intervals. *Perception & Psychophysics, 28*, 493–503.

Schneider, B. (1980b). A technique for the nonmetric analysis of paired comparisons of psychological intervals. *Psychometrika, 45*, 357–372.

Schneider, B., Parker, S., & Stein, D. (1974). The measurement of loudness using direct comparisons of sensory intervals. *Journal of Mathematical Psychology, 11*, 259–273.

Schultz, M., & Joachims, T. (2003). Learning a distance metric from relative comparisons. In S. Thrun,

L. K. Saul, & B. Sch¨olkopf (Eds.), *Advances in Neural Information Processing Systems (NIPS)* (pp. 41–48). Vancouver and Whistler, British Columbia, Canada: MIT Press.

Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika, 27*, 125–140.

Shepard, R. N. (1981). Psychological relations and psychophysical scales: On the status of direct psychophysical measurement. *Journal of Mathematical Psychology, 24*, 21–57.

Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review, 89*, 305.

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology, 48*, 813.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*, 153.

Stevens, S. S. (1961). To honor Fechner and repeal his law. *Science, 133*, 80–86.

Takane, Y. (1978). A maximum likelihood method for nonmetric multidimensional scaling. *Japanese Psychological Research, 20*, 7–17.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., & Kalai, A. (2011). Adaptively learning the crowd kernel. In L. Getoor, & T. Scheffer (Eds.), *International Conference on Machine Learning (ICML)* (pp. 673–680). New York, NY, USA: ACM.

Terada, Y., & von Luxburg, U. (2014). Local ordinal embedding. In E. P. Xing, & T. Jebara (Eds.), *International Conference on Machine Learning (ICML)* (*Vol. 32*, pp. 847–855). Bejing, China: PMLR.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York, NY: John Wiley.

Ukkonen, A., Derakhshan, B., & Heikinheimo, H. (2015). Crowdsourced nonparametric density estimation using relative distances. In *Conference on Human Computation and Crowdsourcing (HCOMP)*.

Van Der Maaten, L., & Weinberger, K. (2012). Stochastic triplet embedding. In *International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6). Santander, Spain: IEEE.

Wichmann, F. A., & Jäkel, F. (2018). Methods in psychophysics. In *The Stevens' handbook of experimental psychology and cognitive neuroscience* (4th ed., Vol. V). Wiley.

Wichmann, F. A., Janssen, D. H. J., Geirhos, R., Aguilar, G., Schütt, H. H., & Maertens, M. et al. (2017). Methods and measurements to compare men against machines. *Electronic Imaging, Human Vision and Electronic Imaging*, *2017*, 36–45.