Eurographics Conference on Visualization (EuroVis) 2018
J. Heer, H. Leitte, and T. Ropinski
(Guest Editors)

Volume 37 (2018), Number 3

Visualizing Multidimensional Data with Order Statistics

M. Raj & R. T. Whitaker

University of Utah, USA

Abstract

Multidimensional data sets are common in many domains, and dimensionality reduction methods that determine a lower dimensional embedding are widely used for visualizing such data sets. This paper presents a novel method to project data onto a lower dimensional space by taking into account the order statistics of the individual data points, which are quantified by their depth or centrality in the overall set. Thus, in addition to conveying relative distances in the data, the proposed method also preserves the order statistics, which are often lost or misrepresented by existing visualization methods. The proposed method entails a modification of the optimization objective of conventional multidimensional scaling (MDS) by introducing a term that penalizes discrepancies between centrality structures in the original space and the embedding. We also introduce two strategies for visualizing lower dimensional embeddings of multidimensional data that takes advantage of the coherent representation of centrality provided by the proposed projection method. We demonstrate the effectiveness of our visualization with comparisons on different kinds of multidimensional data, including categorical and multimodal, from a variety of domains such as botany and health care.

CCS Concepts

•***Human-centered computing*** → ***Information visualization;*** *Visual analytics;* •***Mathematics of computing*** → *Mathematical optimization;*

1. Introduction

Multidimentional data appear frequently in a wide range of domains and applications. For example, data from domains such as healthcare, engineering, and social sciences often contain a large number of dimensions [Lic13]. The various dimensions in such data can contain either numerical or categorical values. Multidimensional data can also have complex structures, for example, the data can be multimodal with several clusters or lie on a lower dimensional manifold in a high dimensional space. A wide range of visualization methods have been developed to help visualize and understand such complex, high-dimensional data sets [LMW*17].

Among the various methods for analyzing high dimensional data, dimensionality reduction methods that project data onto lower dimensional spaces are often useful for getting a quick and general overview of the data. These include various linear and nonlinear methods such as principal component analysis (PCA), multidimensional scaling (MDS), and t-distributed stochastic neighbor embedding (t-SNE). The objective of these methods is often to convey the structure of data by preserving approximate pairwise distances from the original or intrinsic space, in the lower dimensional embedding space. Methods such as PCA and MDS can be formulated to work with *only* inner product information, which is useful for visualizing data in kernel spaces, which may lack an explicit vector representation [SSM97]. Dimensionality reduction techniques are also used in conjunction with other visualization methods [RRT99, MLL12].

Despite usefulness of dimensionality reduction methods for visualizing multidimensional data, there are a few critical limitations associated with those methods. The PCA and related subspace based approaches may not be suitable if the data is not well approximated by a linear subspace. While MDS and nonlinear methods such as t-SNE are able to highlight geometric relationships, even in the presence of nonlinear structure, they are susceptible to misrepresenting the statistical structure in the data. For example, points that are rare and on the outer periphery of a distribution in a high dimensional space may be projected close to a more typical point near the center of the distribution. Such instances are common, and unsurprising if we consider that the objective of those methods is typically to preserve the relative distances between points with no mechanism to correctly convey how central or typical points are in a data set or distribution. While the focus on preserving relative distances to reveal high level structure can be useful, doing so at the expense of centrality information can hinder a true understanding of the data set as a whole, and be particularly detrimental for the purpose of analyzing outliers [Wil17].

In this paper, we propose a novel method to project multidimensional data onto a lower dimensional space while (approximately) preserving centrality structure as well as relative distances

in the data. The focus of this work is different from prior work in robust multidimensional scaling that aim to mitigate the undesirable effects resulting from inconsistencies in data (pairwise distances in the original space) [SL89, FG12]. In contrast, the proposed method is relevant even when there are no inconsistencies in the data. The proposed method does share ideology with a family of methods from the domain of graph drawing, where the goal is to determine node positions in a drawing that simultaneously convey graph-theoretic, internode distances (distances along edges) as well as node centrality or importance based on complementary, graph-theoretical measures [BKW03, BP09, BG14, RW17]. The internode distances in the drawing approximate the graph-theoretical distances under the constraint that the distance of each nodes from the drawing's geometric center be proportional to its graph centrality value.

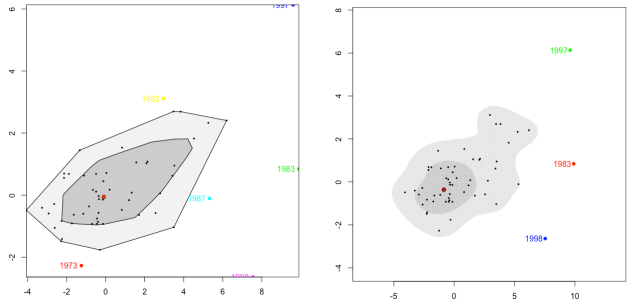
In this work, we aim to preserve *order statistics* of the data in the original space by ensuring that less central or outlying points do not end up appearing to be more central in low dimensional embeddings, or vice versa. We also want to preserve relative pairwise distances in the data as much as possible. An overview of the proposed projection method for satisfying the above objectives is as follows. We first quantify centrality of each member in the original space by employing data depth methods (see Sec 2.1). Next, we design a penalty term to be added to the MDS optimization objective which penalizes low dimensional embeddings, where along any ray traveling away from the position of the most central member, *less central* points are situated further from the center than more central points. Although we demonstrate the proposed method with help of the MDS objective, the general approach can be used to similar effect with any other dimensionality reduction method that involves iterative optimization.

The goal of visualizations, in general, is to highlight features of interest in the data. These feature often include summary statistics such as most central or typical member (also known as the median), least central or outlier members, as well as the shape and the spread of the bulk of data. In case of 1-dimensional (1D) and 2-dimensional (2D) data, visualizations such as the Tukey boxplot [Tuk75] and the bivariate bagplot [RRT99] convey a visual summary of the data by displaying summary statistics. In this paper, we exploit the coherent centrality structure in the embedding space afforded by the proposed projection method to develop visualization strategies, along the lines of the bivariate bagplot, for multidimensional data (i.e. where $d > 2$).

The main contributions of this paper are:

- A novel method for projecting multidimensional data using order statistics called *order aware projection* (OAP).
- Two visualization strategies based on the proposed projection method, namely, *field overlay plot* and *projection bagplot*.
- An interactive prototype tool to explore data.
- Demonstration of the effectiveness of the method with four real data sets.

The rest of the paper is organized as follows. Sec 2 provides an overview of the technical background related to the proposed methods. Sec 3 presents a description of the proposed dimensionality reduction method and visualization strategy. We demonstrate proposed methods using real data in Sec 4, which is followed by a general discussion in Sec 5.



(a) (b)

Figure 1: *(a) Bivariate bagplot and (b) high density region (HDR) boxplot visualizations of El Nino dataset (12-dimensional temperature data for each year from 1951 to 2007) generated using the R Rainbow package [SHS16].*

2. Background

Here we provide an overview of necessary technical background and related work.

2.1. Order Statistics and Data Depth

Order statistics for a data set are members from the data set placed in an ascending order based on some criteria. For our purpose, we are interested in *center-outward* order statistics that help quantify how central or outlying a member is with respect to a data set. In the case of 1D numeric data, sorting numbers based on distance from the median provides an easy way to obtain order statistics. When the data is multidimensional, a family of methods from descriptive statistics known as *data depth* can be used to quantify center-outwardness. Data depth methods exhibit several useful properties, which make it an attractive basis for analyzing data. These properties include robustness, maximum at center, monotonicity, and zero at infinity [ZS00].

Data depth methods have been proposed for tackling several types of multidimensional and multivariate data, for example, high-dimensional points [Tuk75], functions [LPR09], sets [WMK13], multivariate curves [MWK14], and paths on a graph [RMR*17]. In this paper, we use different formulations of data depth based on the type of data. We use halfspace depth for numerical multi-dimensional data with relatively few dimensions (Sec 4.2). We use functional depth for dealing with higher dimensional data because it can be efficiently computed for such data (Sec 4.1). Finally, we use set depth for categorical data sets (Secs 4.3 and 4.4).

A brief overview of halfspace depth, functional depth, and set depth follows. Halfspace depth of any point $\mathbf{x} \in \mathbb{R}^d$ with respect to a set of points $X \in \mathbb{R}^d$ is defined as the smallest number of data

points from X that can be contained in a closed half space also containing $\mathbf{x}$ [Tuk75, DM16]. This can be stated as:

$$d_{\text{halfspace}}(\mathbf{x}|X) = \min_{\mathbf{a}\in\mathbb{R}^d\setminus 0} \left|\left\{\mathbf{p}\in X \;:\; \langle \mathbf{a},\mathbf{p}\rangle \geq \langle \mathbf{a},\mathbf{x}\rangle\right\}\right| \qquad (1)$$

Functional depth of any function $g(t)$ with respect to a set of functions $F = \{f_i(t) : 1 \leq i \leq n\}$, $f_i : \mathcal{D} \to \mathcal{R}$, where $\mathcal{D}$ and $\mathcal{R}$ are intervals in $\mathbb{R}$, is given by the probability of $g(t)$ being contained in a functional band, where functional band is the region between the min/max envelope formed by a set of j randomly chosen functions $\{f_1(t),\ldots,f_j(t)\} \in F$ [LPR09]. This can be stated as:

$$d_{\text{functional}}\left(g(t)|F\right) = \text{Prob}\left(g(t) \subset \text{fB}\left[\{f_1(t),\cdots,f_j(t)\}\right]\right), \qquad (2)$$

where $\text{fB}[\cdot]$ denotes the functional band. A function $g(t)$ is contained in the functional band formed by $\{f_1(t),\ldots,f_j(t)\}$ if it satisfies the following:

$$\begin{aligned} g(t) \subset \; &\text{fB}\left[\{f_1(t),\cdots,f_j(t)\}\right] \quad \text{iff} \\ &\min\left(f_1(t),\cdots,f_j(t)\right) \leq g(t) \leq \max\left(f_1(t),\cdots,f_j(t)\right) \quad \forall t. \end{aligned} \qquad (3)$$

Set depth of any set s with respect to a set of sets $S = \{s_i : 1 \leq i \leq n\}$ is given by the probability of s being contained in a *set band*, where set band is the set bounded by the union and intersection of j randomly chosen sets $\{s_1,\ldots,s_j\} \in S$ [WMK13]. This can be stated as:

$$d_{\text{set}}\left(s|S\right) = \text{Prob}\left(s \subset \text{sB}\left[\{s_1,\ldots,s_j\}\right]\right), \qquad (4)$$

where $\text{sB}[\cdot]$ denotes the *set band*. A set s is contained in the set band formed by $\{s_1,\ldots,s_j\}$ if it satisfies the following:

$$s \subset \text{sB}\left[\{s_1,\ldots,s_j\}\right] \quad \text{iff} \quad \bigcup_{k=1}^{j} s_k \subset s \subset \bigcap_{k=1}^{j} s_k.$$

Function depth and set depth are stable with respect to the choice of j where $2 \leq j \leq n$ [LPR09, WMK13].

2.2. Data Depth based Visualizations

A common area of application for data depth methods is ensemble visualization where the order statistics obtained using data depth are used to design summary visualizations for ensembles of various kinds of data. The perhaps most well known example is the Tukey boxplot [Tuk75]. Other depth based visualizations have been proposed for bivariate data [RRT99], high-dimensional data [Hyn96], ensembles of functions [SG11], surfaces [GJP*14], sets or isocontours [WMK13], curves [MWK14, LPSLG14], and paths on graphs [RMR*17]. Our work relates closely to the visualizations for multidimensional data, particularly the bivariate bagplot [RRT99] and the high density region (HDR) boxplot [HS10] (see Fig 1).

For 2D data, the bivariate bagplot (Fig 1a) is a visualization technique that highlights the median, spread, skewness, and outliers in the data. The first step for drawing a bagplot is to determine order statistics using half space depth. This is followed by drawing the inner and outer convex polygons or *bands*. The inner band highlights the most central half of the data as determined by the order statistics while the outer band is constructed by inflating the inner band by a constant factor α. Points outside the outer band are considered to be outliers. The HDR boxplot uses bivariate kernel density estimation to identify regions of interest. The bivariate bagplot as well as the HDR boxplot use dimensionality reduction methods, typically PCA, for dealing with higher dimensional data ($d > 2$) by projecting the data to 2D as a preprocessing step [Hyn96, HS10].

2.3. Multidimensional Scaling (MDS)

Since the proposed projection method uses the MDS objective function, we give a brief overview of MDS and its usage in dimensionality reduction. MDS refers to a popular class of techniques for visualizing similarities between members of a data set. Given a collection of high-dimensional points $Y = [\mathbf{y}_1,\ldots,\mathbf{y}_n]^T \in \mathbb{R}^{n\times d}$, the goal of MDS is to find a low-dimensional embedding $X = [\mathbf{x}_1,\ldots,\mathbf{x}_n]^T \in \mathbb{R}^{n\times k}$, where $k < d$, such that the discrepancy between the pairwise distances in the original space $\mathbb{R}^d$ and corresponding distances in the embedding space $\mathbb{R}^k$ is minimal.

While there are several variants of MDS, in this paper we use a variant known as metric MDS with distance scaling; without loss of generality. Distance scaling makes this variant of MDS nonlinear with more emphasis on conveying smaller distances. The objective function of metric MDS is also known as *stress*, and after incorporating distance scaling, it can be written as follows [BG05, McG66]:

$$\sigma(X) = \sum_{i<j} w_{ij}\left(\delta_{ij} - d(\mathbf{x}_i,\mathbf{x}_j)\right)^2, \qquad (5)$$

where $\delta_{ij} = ||\mathbf{y}_i - \mathbf{y}_j||_2$, $d(\mathbf{x}_i,\mathbf{x}_j) = ||\mathbf{x}_i - \mathbf{x}_j||_2$, and $w_{ij} = \delta_{ij}^{-2}$. The gradient of the above MDS objective can be written as follows [BG05]:

$$\nabla\sigma(X) = 2VX - B(X)X \qquad (6)$$

where matrices $V = (v_{ij})$ and $B = (b_{ij})$, with $1 \leq i,j \leq n$, can be represented as:

$$v_{ij} = \begin{cases} -w_{ij} & \text{for } i \neq j \\ \sum_{j=1,j\neq i}^{n} w_{ij} & \text{for } i = j \end{cases} \qquad b_{ii} = -\sum_{j=1,j\neq i}^{n} b_{ij}$$

$$b_{ij} = \begin{cases} -\frac{w_{ij}\delta_{ij}}{d(\mathbf{x}_i,\mathbf{x}_j)} & \text{for } i \neq j \text{ and } d(\mathbf{x}_i,\mathbf{x}_j) \neq 0 \\ 0 & \text{for } i \neq j \text{ and } d(\mathbf{x}_i,\mathbf{x}_j) = 0 \end{cases}$$

2.4. Monotone Regression along One Variable for Multivariate Data

The proposed projection method also involves computing a continuous and smooth, radially decreasing approximation of depth of members in a 2D embedding. We call this approximation the *monotonic field* (Fig 2c). Note that data depth values are computed for points in the original space and not after they are projected onto the embedding space. To construct a monotonic depth field from a

sparse set of depth values arranged in a 2D embedding plane, we start by computing a smooth *interpolated field* (Fig 2b) of depth values using the thin plate spline technique [Boo89]. In what follows, we briefly describe our approach for computing the monotonic field by radially monotonizing the interpolation field. This approach is adapted from a technique for performing monotone regression for multivariate data [DS06].

The process of computing radially monotonic approximations of a smooth 2D field depends on a method to find monotonic approximations of univariate data. Given a smooth 1D function $m(t) : [0,1] \to \mathbb{R}$, the following two steps provide a monotonic approximation, $\hat{m}_A(t)$, that is smooth and first-order asymptotically equivalent to $m(t)$ [DNP*06]:

- Step 1 (monotonization): Sample input function at regular intervals, compute a density estimate of the samples, and then compute a cdf of the density estimate to arrive at the inverse of the monotonic approximation.

$$\hat{m}_A^{-1}(z) = \frac{1}{N\omega} \sum_{i=1}^{N} \int_z^{\infty} K\left(\frac{m\left(\frac{i}{N}\right) - u}{\omega} \right) du \tag{7}$$

  where N controls the sampling resolution, and K is a smooth, symmetric kernel with bandwidth ω.
- Step 2 (inversion): Calculate the inverse of $\hat{m}_A^{-1}$, which is the desired monotonically decreasing approximation of the 1D function $m(t)$.

For computing a radially monotonic approximation of the interpolated field, we proceed by resampling the field onto a polar grid centered at the median (deepest member as per data depth computed in the original space). We then treat values on the field along each of the evenly spaced angular coordinates as 1D functions, which can then be monotonized using the procedure described above. On monotonizing those 1D functions along each direction, we arrive at the monotonic field, which we then resample back to Cartesian coordinates. The resolution of the polar grid, both radial and angular, determine the quality (smoothness) of monotonic field (we use 360 radial divisions for results in this paper). The smoothness of the interpolation field is preserved through this process, meaning that field values along adjacent directions vary smoothly and remain coherent, due to the properties of the monotonization process (except at the origin due to the intermediate polar coordinate representation) [DS06].

3. Method

Here we describe the proposed projection method, which preserves the centrality structures using order statistics (Sec 3.1), and visualization strategies, which use the resultant embedding (Sec 3.2).

3.1. Projecting Multidimensional Data using Order Statistics (Order Aware Projection)

The high-level goal of our projection method is to preserve both the relative distances between individual members as well as the order statistics from the original multidimensional space when computing a lower dimensional embedding. To achieve this, we design an objective function which comprises of two terms. The first term

Algorithm 1: Order Aware Projection (OAP)

Input: $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T \in \mathbb{R}^{n \times d}$, maximum number of iterations $i_{\max} \in \mathbb{N}$, depth field lag ℓ, step size τ, depth weight w_p

Output: Positions $X_{i_{\max+1}} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times k}$ where $k < d$

$X_0 \leftarrow$ compute initial embedding using MDS ; /* (2.3) */

$\mathbf{h} \in \mathbb{R}^{n \times 1} \leftarrow$ compute order statistics for $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \in \mathbb{R}^d$

for $i = 1, \ldots, i_{max}$ **do**

  if $i \mod \ell = 1$ **then**

    $X' \leftarrow X_i$

    $M_{X',\mathbf{h}}(X_i) \leftarrow$ compute monotonic field ; /* (2.4) */

  end

  $X_{i+1} \leftarrow$

  $X_i - \tau\Big(\nabla\sigma(X_i) + w_p \times 2\big(M_{X',\mathbf{h}}(X_i) - \mathbf{h}\big) \odot \nabla M_{X',\mathbf{h}}(X_i)\Big)$;

  /* perform gradient update (3.1) */

end

is identical to the MDS stress (Sec 2.3), which penalizes discord in pairwise distances between intrinsic space and the embedding. The second term levies an energy penalty for discord in the center-outward order statistics. Since the order statistics are determined using data depth, we call this term the *depth penalty*.

The data depth values computed in the original space (Sec 2.1) and the monotonic field computed in the embedding space (Sec 2.4) are used to quantify the discord in centrality structure between the original and embedding spaces. The isocontours of the monotonic field mimic the monotonic, center-outward decrease of depth values in the original space. The depth penalty at each point is proportional to the difference between its depth value in the original space and the depth values of the monotonic field sampled at the location of its projection in the embedding space, and can be expressed as follows:

$$p(X) \propto \big(M_{X,\mathbf{h}}(X) - \mathbf{h}\big)^2 \tag{8}$$

where $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times k}$, $\mathbf{h} \in \mathbb{R}^{n \times 1}$ contains depth values associated with X computed in the original space $\mathbb{R}^d$, where $k < d$, and $M_{X,\mathbf{h}}(X) \in \mathbb{R}^{n \times 1}$ denotes values of the 2D monotonic field at positions in X. The mention of X and $\mathbf{h}$ in the subscript indicates their use in the construction of the monotonic field, while X in parenthesis indicates positions where field values are sampled. If the interpolated field (Sec 2.4) is also itself radially monotonic, the value of depth penalty term approaches zero. The complete objective function, which includes both MDS stress and depth penalty, can be stated as follows:

$$\gamma(X) = \underbrace{\sigma(X)}_{\text{MDS stress}} + \; w_p \, p(X) \tag{9}$$

where w_p is a constant of proportionality controlling the relative importance of the depth penalty with respect to MDS stress. The

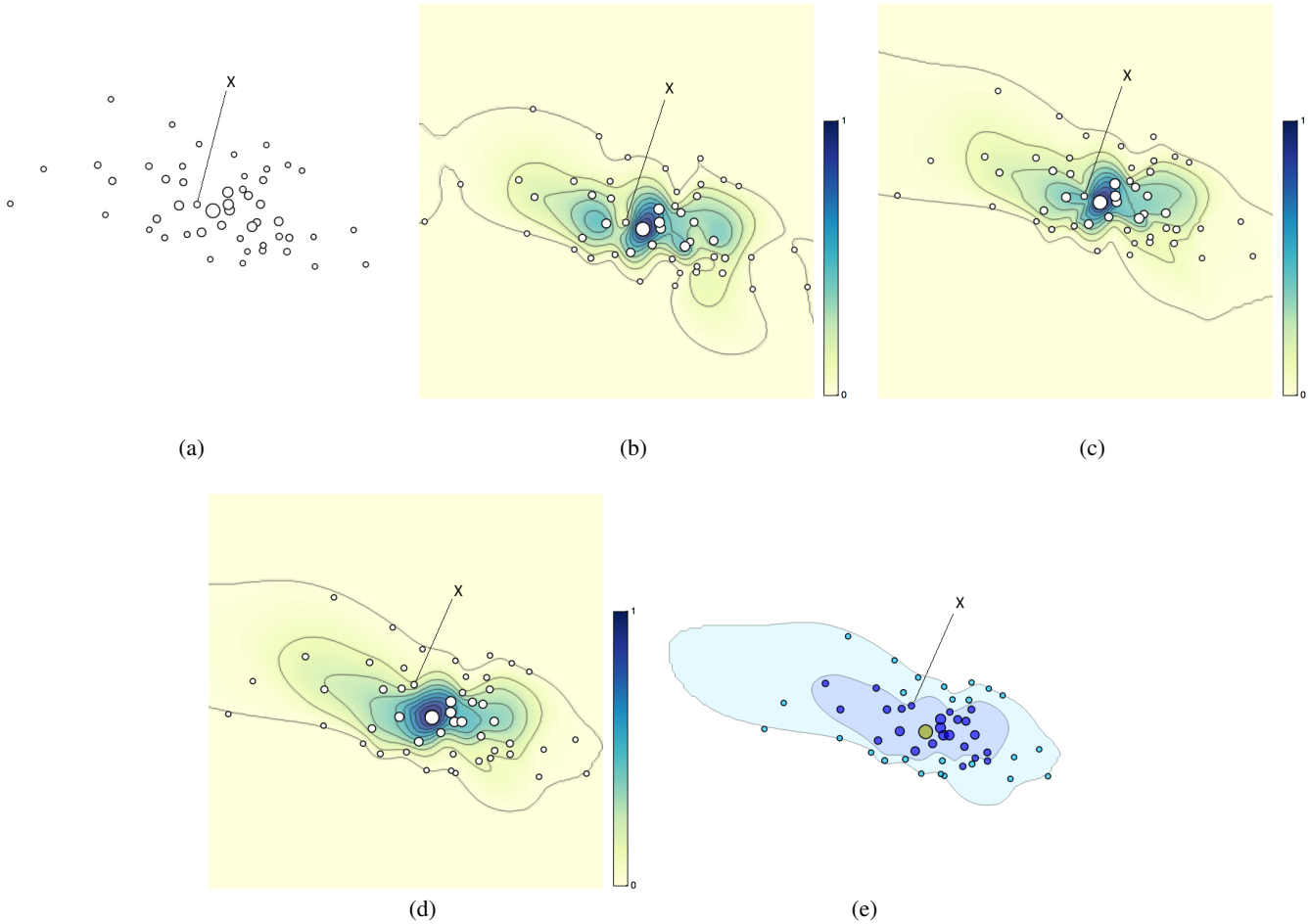


Figure 2: *Various stages during the proposed methods. a) Points from an anisotropic, 3D normal distribution projected on a 2D plane using MDS. Circle sizes indicate half space depth of points in the original 3D space. b) The initial interpolated field in the background of the MDS projection. c) The initial monotonic field in background obtained from initial interpolated field. d) Field overlay plot using order aware projection (OAP) after optimization is complete. The final monotonic field shown in the background. e) Projection bagplot visualization. Median is shown in yellow. Deep blue indicates 50 percent band and light blue indicates 100 percent band.*

gradient of the above objective can be derived as:

$$\nabla\gamma(X) = \nabla\sigma(X) + w_p \times \underbrace{2\big(M_{X,\mathbf{h}}(X) - \mathbf{h}\big) \odot \nabla M_{X,\mathbf{h}}(X)}_{\nabla p(X)} \quad (10)$$

where $\odot$ denotes element-wise product. We perform optimization of the above objective using gradient descent until X converges or the maximum number of allowed iterations is reached. The proposed projection method is summarized in Algorithm 1.

Optimization of the above objective requires a few considerations in practice. First, computing the gradient of the monotonic field M at positions X is nontrivial due to dependence of M itself on X. We deal with this issue by letting the field lag, which means to recompute M only after a fixed number of iterations, ℓ, have passed since the previous update and treat it as a constant during all intervening iterations (see Fig 3). If field M is help constant ($\ell = \infty$), convergence at a local minima can be guaranteed due to properties of gradient descent. On allowing field to lag suitably ($1 \leq \ell < \infty$, see Sec 5), in practice we observe convergence to a lower energy state; although a theoretical guarantee remains a topic for future work. This approach is also used for minimizing similar energies for computing graph layouts [RW17]. Second, for stability with regard to the median, the proposed method relies on known robustness of data depth methods in situations of data contamination [Tuk75, SG11, WMK13]. In cases where there are multiple members identified as a median in the original space, we choose the member with the highest depth value among them in the embedding space. This helps reduce overall energy if different medians are projected far apart, as is often observed with categorical data (Secs 4.3 and 4.4).

In the case of multimodal data where class membership information is known in advance, we construct a separate monotonic field for each class centered at the median of that class, which leads to separate, exclusive depth penalty terms that apply only to the members of the associated class (Secs 4.2 and 4.4). The MDS

term is the same as in the general case (which assumes a unimodal distribution) and considers pairwise distance relationships across the entire data set. This approach preserves the centrality structure within each class while allowing MDS forces to determine the relative placement of different classes.

3.2. Field Overlay and Projection Bagplot Visualizations

At the end of the optimization process, all members in the data set are aligned with their corresponding isocontours (whose isovalue matches member depth) on the underlying monotonic field. The shape of the isocontours depends on the data and can often provide useful insights into the structure of the data in the original space. Our first visualization strategy, called *field overlay plot*, is to present the order aware projection (OAP) embedding overlaid on the associated monotonic field (Fig 2d). We show the monotonic field as a color heatmap with isocontour lines for 10 equidistant values spanning the range of depth values. This approach helps with the interpretability of the OAP embedding by highlighting the depth associated with each member as well as the regions/directions of fast and slow depth changes.

Due to the radially, monotonically decreasing property of the monotonic field, all isocontours divide the embedding space into inner and outer regions, which exclusively contain members with higher and and lower depth in the original space. We propose the *projection bagplot* visualization (Fig 2e), which uses this arrangement of members in the embedding space to convey the median, inner, and outer bands analogous to those seen in the Tukey boxplot [Tuk75]. The depth value of the isocontour corresponding to the 50 percent band, $h_{50\%}$, is chosen to be the value of the member at 50th percentile by ranking the members' depth values. The 100 percent band is formed by inflating the 50 percent band by a constant factor α. So we have $h_{100\%} = h_{\text{median}} - \alpha \times (h_{\text{median}} - h_{50\%})$ where h_{median} is the depth value of the median (highest depth value by definition) and $\alpha = 1.5$ typically [Tuk75, SG11]. We use higher and lower color saturation for indicating band/members in the 50 percent and 100 percent bands, respectively. For multimodal data with known class membership information, a separate set of 50 percent and 100 percent bands is computed and displayed for each class (Figs 6 and 8).

The projection bagplot visualization shares some similarities with both the bagplot and the HDR bagplot. The interpretation of the bands in the proposed method is similar to that for the bagplot, while the shape of bands is smooth and star shaped like in the HDR bagplot. Despite the similarities, the proposed method is notability different in its handling of multidimensional data projected to lower dimensions due to the emphasis on maintaining the center-outward order of members during the order aware projection process. While glyph sizes or color can be modulated to convey order statistics, the reliance of bagplot and HDR bagplot visualization on existing dimensionality reduction techniques [HS10] can lead to conflict between glyph size/color and location cues (e.g., members appearing as outliers due to smaller glyphs also appearing closer to the center). Furthermore, when displaying large data sets, available display space may place an upper bound on the glyph sizes, thereby restricting the usable range of glyph sizes.

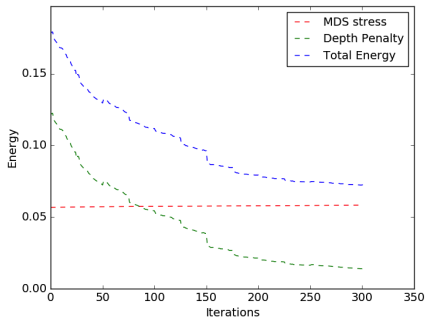


Figure 3: *The typical profile for MDS stress and depth penalty during the optimization process. MDS stress increases slightly. The depth penalty undergoes sharp drops periodically at iterations with monotonic field updates.*

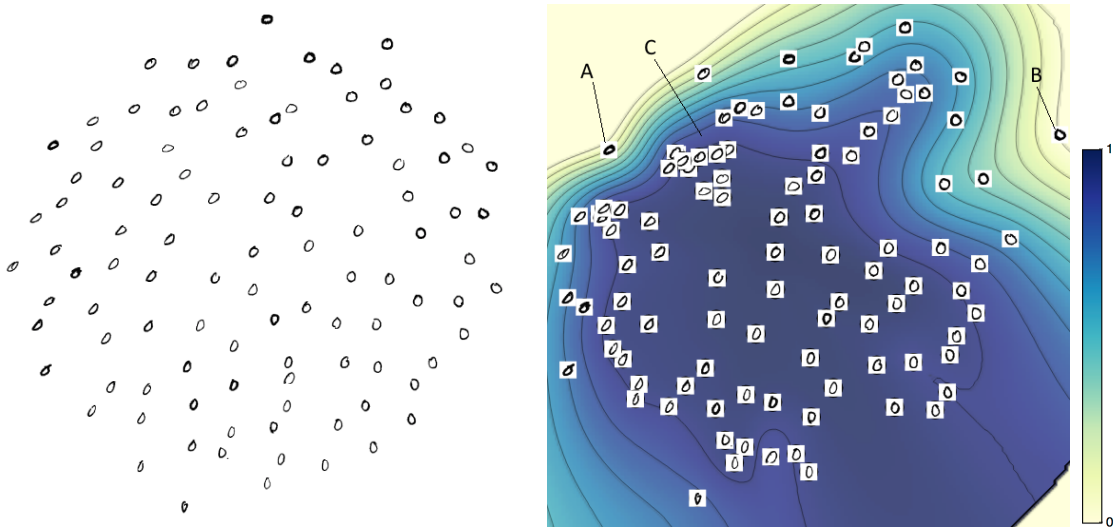
4. Results

We now present some example visualizations of real data sets with existing and proposed methods.

4.1. MNIST Data Set

The MNIST data set is popular in the machine learning community and is comprised of thousands of samples of handwritten digits [LC10]. The samples are formatted as 28×28 pixel gray scale images, resulting in each sample being comprised of 784 dimensions. Fig 4 shows two visualizations of a random subset of 100 samples of digit 0, while Fig 5 shows digits 0, 1, 7 with 100 samples each. We consider each sample to be an instance of a 784-dimensional function and use functional depth to compute order statistics. In Fig 5, we use order statistics computed for each digit separately. These order statistics are used to obtain an OAP embedding, which is used to draw a field overlay plot (Sec 3.2). In Fig 5, we use the proposed visualization strategy for multimodal data with a separate monotonic field for each digit (Sec 3).

On comparing the MDS embedding Fig 4a and the proposed field overlay plot Fig 4b, we can make a few interesting observations. First, we notice that the underlying depth contours make it easy to spot outliers in the field overlay plot. Since the contours adapt to the data, we also notice different outlier characteristics, such as sharing some similarity with other members (see member A) or being more peculiar (see member B). We also notice that the proposed projection (OAP) presents more clique-like structures, often with similar members in a tighter cluster than in the MDS (see cliques around region C). The formation of cliques can be understood by considering that members in a relatively local region of the original space would tend to have similar depth values and low pairwise distances, and would be encouraged to be placed similarly in the embedding by both depth and MDS energies. In Fig 5, we can observe the different monotonic fields for digits 0,1, and 7. The higher overlap of digit 7 with other digits, particularly with digit 1, is immediately clear. Furthermore, on tracing the outermost isocontours of fields, we are immediately drawn towards outlying members that have been placed far from other members, or share similarities with



(a) (b)

Figure 4: *MNIST data sample visualizations: (a) MDS, and (b) field overlay plot using order aware projection (OAP). Outliers and cliques appear more prominent in the field overlay plot.*

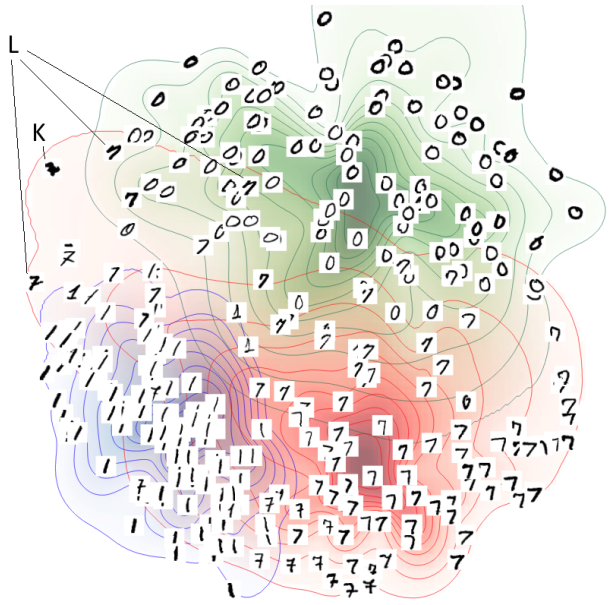


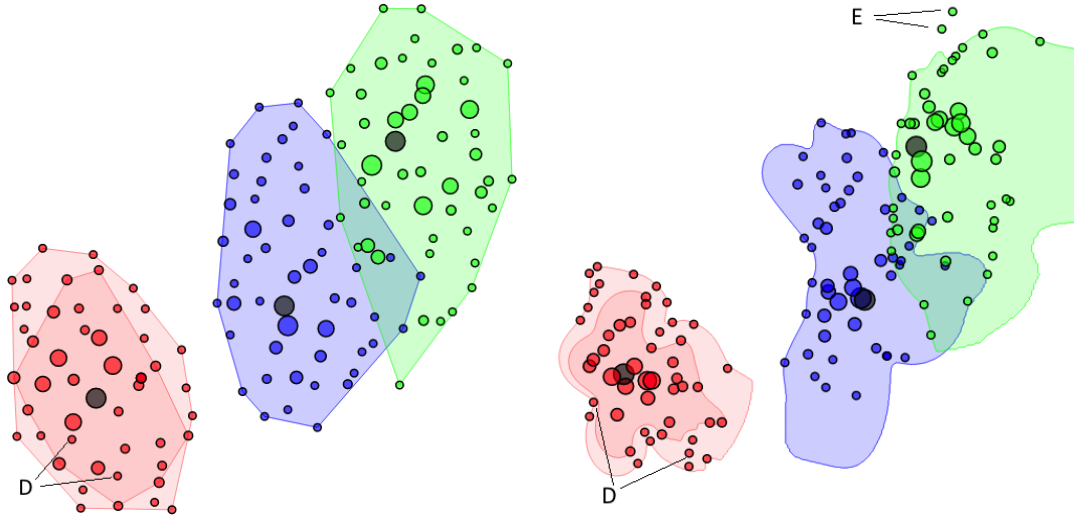
Figure 5: *MNIST data sample visualization for multiple digits (0, 1, and 7) with field overlay plot using order aware projection (OAP). Monotonic fields corresponding to 0,1, and 7, are shown using heatmaps and isocontour lines drawn in green, blue, and red, respectively. Higher saturation of colors in the heatmaps indicate higher value of monotonic field. Unusual members are apparent on tracing outermost isocontours.*

other digits. For example, see instances in marked by K and L, respectively, in Fig 5.

4.2. Iris Flower Data Set

We obtained the well-known Iris data set from the UCI machine learning repository [Lic13]. The data set contains flower sepal and petal measurements from three related species of Iris flowers, and includes 50 instances of each species with four numeric measurements per instance. In Fig 6, we use the proposed visualization strategy for multimodal data with a separate monotonic field for each of the three species classes (Sec 3). The order statistics are computed using half space depth for each class separately. The median of each class is colored dark gray, and the size of circular glyphs encodes the depth of members with respect to their respective classes.

Fig 6a shows a bivariate bagplot [HS10] and Fig 6b shows the proposed projection bagplot. In both figures, we immediately notice a difference in the structure of the classes based on the overlap of the 50 and 100 percent bands. In the red (Setosa) class, we observe a partial overlap as opposed to a full overlap in the blue (Versicolor) and green (Virginica) classes. This indicates a more even spread of members in the red class and more members at the class boundaries for the blue and green classes. We also see that the centrality structures within classes is preserved in the projection bagplot; along all outward directions from the median, the depth of the members falls monotonically. The preservation of centrality structures prevents cases as in region D where members in the 100 percent band are projected to fall inside the 50 percent bands in the embedding in the bivariate bagplot. Another interesting area is region E where two members are pushed out of the 100 percent band despite being of similar depth as other nearby points. This happens due to the distance-preserving aspect of the proposed objective (Eqn 9) trying to convey differences among members that are all on the boundary of the green class. Such cases as highlighted by the projection bagplot are good instances for further exploration.



(a) (b)

Figure 6: *Iris flower data visualization: (a) bivariate bagplot using MDS, and (b) projection bagplot using order aware projection (OAP). There are three species of flowers, each represented by a color, and each circle represents an individual flower. For the blue and green classes, 50 percent and 100 percent bands overlap due to a large proportion of members with identical, lowest value of depth. For the red class, only the projection bagplot conveys band associations of the flowers correctly.*

4.3. Unidentified Flying Object (UFO) Encounters Data Set

We now look at a data set related to UFO encounters that was compiled by Winner from information available in the public domain [Win04]. The data set contains the following six attributes (one numeric and five categorical): year of sighting, location, presence/absence of physical effects, multimedia, extraterrestrial contact, and involvement of abduction. To understand the typical/atypical characteristics of recored UFO encounters across the years, we exclude the year information, and include only the categorical dimensions in our analysis. The distances between members needed for the MDS term are obtained through the inner products computed using the "k0" kernel for categorical data [BMV13]. We compute order statistics for categorical data by using set band depth (Sec 2.1) and treating each member as a set of its attribute values from all dimensions [MWK17].

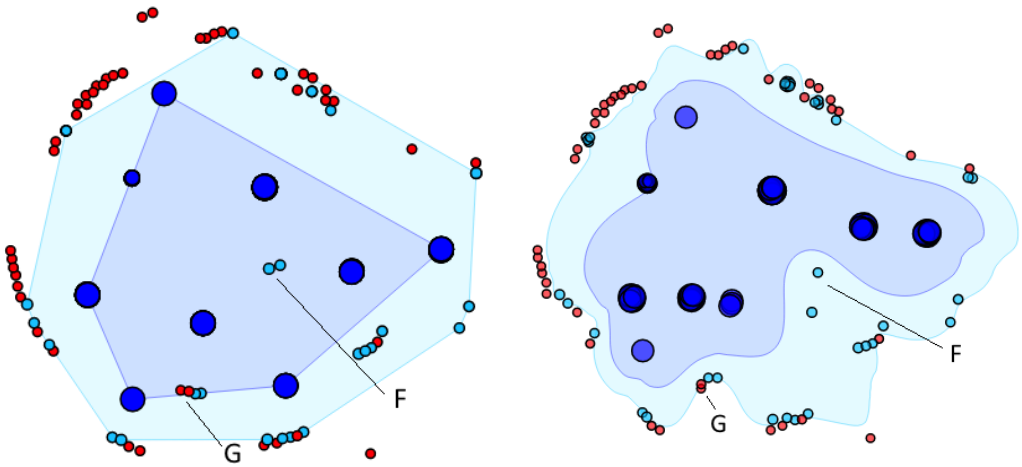
Fig 7 shows two visualizations for the UFO data set. An interesting feature of this data set is the presence of several members with the highest depth value that are placed relatively far from each other. On inspection we find that they are all sightings in the USA, which leads to the conclusion that a large number and variety of UFO sightings are recorded in the USA. Some sightings at other locations share many attributes of US sightings, still cannot be representative of the data, as indicated by their low depth values, because of being at a different location (see region F). The projection bagplot (Fig 7b) is able to convey this well by adjusting the shape of the 50 percent band to exclude those points without significant change in their positions, while the bivariate bagplot (Fig 7a) shows a contradiction where members supposed to be in the 100 percent band are seen within the 50 percent band. Another such contradiction is seen in region G where outliers (shown in red) appear to be inside the 100 percent band.

4.4. Breast Cancer Data Set

Fig 8 displays our final data set, which consists of a collection of breast cancer patient attributes compiled at the University Medical Center at Ljubljana and made available by the UCI machine learning repository [ZS88, Lic13]. This data set contains two patient classes, recurrence and nonrecurrence, with 85 and 201 instances per class, respectively. There are nine attributes per instance such age range, tumor size, degree of malignancy, etc. Analogous to the approach in Sec 4.3, we use a categorical kernel to compute distances in the original space [BMV13]. Since the data is bimodal with known class membership information, we use the proposed projection and visualization strategy for multimodal data with a separate monotonic field for each class (Sec 3). The two medians are drawn as larger circles in the color of their respective class.

This is a case of bimodal data where the classes are not clearly separated. We notice from Fig 8 that the nonrecurrence class is somewhat coherent while the recurrence class is more spread out in a ring-like distribution. Such a distribution is a case that highlights the distinction between data depth and data density. While depth would be high at the geometric center of such a distribution, density would be low due to absence of members near the center. Form this distribution, we can infer that there must be a large variation among the member attributes of the recurrence class, with no good options among members to be considered typical or most representative.

As expected, the projection bagplot visualization has members in both classes arranged radially, in order of decreasing depth from their respective class medians. The resulting structure makes it easier to spot several interesting outlying cliques. For example, instances near region H correspond to relatively younger individuals



(a) (b)

Figure 7: *Unidentified flying object (UFO) encounters data visualizations: (a) bivariate bagplot using MDS, and (b) projection bagplot using order aware projection (OAP). Each circle represents an encounter. Deep blue, light blue and red circle colors indicate association to the 50 percent band, 100 percent band, and outliers. The projection bagplot is able to show correct band associations for members while the bivariate bagplot misplaces some encounter instances with respect to bands.*

with moderate to high tumor size and malignancy and appear to be outliers with respect to both recurrence and nonrecurrence classes. Another interesting region of interest is J where there are instances of older individuals with large tumor size and varying malignancy also appearing as outliers with regard to both classes.

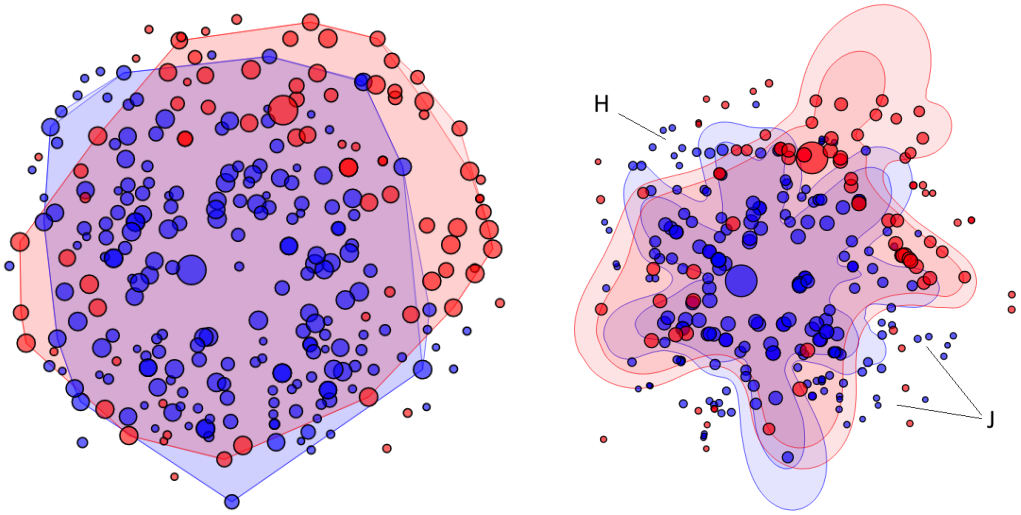
5. Discussion

This paper provides a solution to visualize high-dimensional data ($d \geq 3$) with order statistics in a manner that is popular for visualizing lower dimensional datasets. Similar members are positioned close in the embedding, and central or typical members appear to be more toward the center than outlying or atypical members. To achieve such an embedding, one might consider simply augmenting the data with an additional dimension containing data depth values. However, such an approach fails to take into account the anisotropic structure of data, and pushes for members with similar depth to be placed at similar distance from center regardless of direction from center; with members having higher depth value being placed closer to center. On the other hand, OAP allows more flexibility by allowing the rate of fall of depth values to vary smoothly across adjacent directions, as long as depth values drop monotonically along each direction. This flexibility is helpful to better preserve pairwise distances, particularly in frequently seen cases where points near the boundary along the minor axis are projected close to the median. For example, the proposed methods allow the point X in Fig 2 to remain close to the median while also indicating that it is more outlying than it appears in the MDS projection (compare Fig 2a versus Figs 2d and 2e).

Oftentimes, multidimensional data to be analyzed is heterogeneous, which means that it includes both numerical and categorical dimensions. Our approach is able to handle such data since its only requirement is a way to compute distances and center outward order statistics, which can be computed for such data [MWK17]. Such data are often seen as input for machine learning tasks (e.g., classification or clustering), and the proposed methods can be valuable to understand the structure of kernel spaces where those tasks are performed. A key feature of the proposed projection method (OAP) is its ability to integrate distances and order statistics from different spaces as shown in Secs 4.3 and 4.4. We may also use order statistics with any other (possibly non data depth) method that may be appropriate for the application at hand [Wil17].

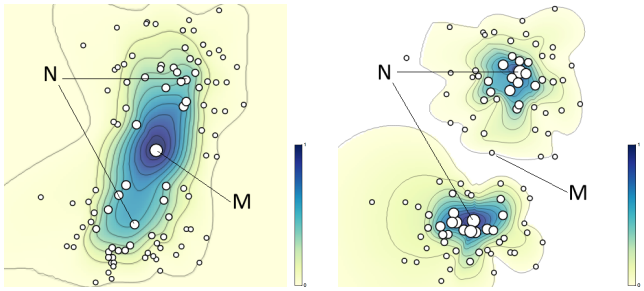
In case of multimodal data *without* known class membership information, OAP can lead to significant misrepresentation of distances if we compute data depth with respect to all points. This is because standard data depth methods measure geometric centrality could assign high depth values for points in region between the clusters, even if the region is sparsely populated; low density does not imply low centrality (see point M in Fig 9a). Furthermore, geometric centers of various clusters would be assigned low depth values if they do not lie near the geometric center of the entire distribution (see points N in Fig 9a). Such an assignment of depth, although technically correct , can lead to an embedding where cluster centers seem less prominent than surrounding points. One way to make cluster centers more prominent, which may be important in multimodal data, is to first cluster the data and then compute data depth, and monotonic fields, for each cluster separately (see Fig 9b).

The proposed projection method (OAP) requires manual adjustment of two parameters: ω_p, which controls the relative emphasis on the centrality structure with respect to preserving pairwise distances, and ℓ, which controls the lag between updates of the monotonic field. Too small values of ω_p will converge to the MDS layout, while too large values of ω_p can cause unnecessary distor-



(a) (b)

Figure 8: *Breast cancer data visualizations: (a) bivariate bagplot using MDS, and (b) projection bagplot using order aware projection (OAP). Each circle represents a set of patient attributes. The data contains two classes based on patient outcomes: recurrence (red) and nonrecurrence (blue). The recurrence class is seen to deviate from normal while the nonrecurrence class presents a more coherent distribution based on recorded attributes.*



(a) (b)

Figure 9: *Field overlay plots using order aware projection (OAP) for synthetically generated 3D multimodal data set with unknown class membership. Order statistics are computed using half space depth (a) for all points in data set together, and (b) for each cluster separately after a clustering step using k-means method.*

tion. We find that values between 1 and 3 for ω_p provide a good balance. In case of ℓ, too small can cause instability that prevents convergence. The instability arises due to a possibility of (typically small) increase in overall energy accompanying computation of the monotonic field. With sufficiently large ℓ, this increase is more than compensated after points adjust to the new field. On the other hand, too large values of ℓ can delay convergence due to delayed spline updates. We use $\ell = 25$ for all examples in this paper. During the iterative optimization process, the computational cost of iterations involving an update of the monotonic field is $\mathcal{O}(n^3)$—arising from computation of the thin plate spline (Sec 2.4). However, the majority of iterations do not involve field updates and incur a lower cost of $\mathcal{O}(n^2)$ operations.

6. Future Work

An important area of application for our method is the visualization of data in kernel spaces (Secs 4.3 and 4.4). While we use set band depth for kernel-based examples in this paper to obtain order statistics, often the only option is to compute depth directly in the kernel space, for example, in the case of ensembles of structured data such as chemical compound graphs [DLdCD*91]. Since existing methods for computing depth are not suitable for *high-dimensional kernel spaces*, which is often the case with graph kernels [VSKB10], a method to compute depth in such spaces would expand the scope of data that could be visualized using the proposed method. Such a method to compute depth would need to address the limitations of existing methods by being efficiently computable in high- dimensional spaces as well as having an inner product based formulation for operating in kernel spaces.

Another exciting avenue for future work would be to extend the proposed approach to work with manifold-based dimensionality reduction techniques such as Isomap and tSNE, which motivates the need to develop data depth methods that are also able to operate with respect to manifolds. Automatic estimation of parameter values based on the data to achieve an optimum balance between conveying distances and centrality would also be useful. Finally, projection bagplot visualization could complement other methods for set visualization such as tabplot [TdJD*13] and parallel coordinates [YNMX17] as part of an integrated, interactive system with linked views.

7. Acknowledgments

This material is based upon work supported by the National Science Foundation under grant IIS-1212806.

References

[BG05] BORG I., GROENEN P. J.: *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005. 3

[BG14] BAINGANA B., GIANNAKIS G. B.: Embedding graphs under centrality constraints for network visualization. *arXiv preprint arXiv:1401.4408* (2014). 2

[BKW03] BRANDES U., KENIS P., WAGNER D.: Communicating centrality in policy network drawings. *IEEE transactions on visualization and computer graphics 9*, 2 (2003), 241–253. 2

[BMV13] BELANCHE MUÑOZ L. A., VILLEGAS M.: Kernel functions for categorical variables with application to problems in the life sciences. In *Artificial intelligence research and development: proceedings of the 16 International Conference of the Catalan Association of Artificial Intelligence* (2013), pp. 171–180. 8

[Boo89] BOOKSTEIN F. L.: Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence 11*, 6 (1989), 567–585. 4

[BP09] BRANDES U., PICH C.: More flexible radial layout. In *International Symposium on Graph Drawing* (2009), Springer, pp. 107–118. 2

[DLdCD*91] DEBNATH A. K., LOPEZ DE COMPADRE R. L., DEBNATH G., SHUSTERMAN A. J., HANSCH C.: Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry 34*, 2 (1991), 786–797. 10

[DM16] DYCKERHOFF R., MOZHAROVSKYI P.: Exact computation of the halfspace depth. *Computational Statistics & Data Analysis 98* (2016), 19–30. 3

[DNP*06] DETTE H., NEUMEYER N., PILZ K. F., ET AL.: A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli 12*, 3 (2006), 469–490. 4

[DS06] DETTE H., SCHEDER R.: Strictly monotone and smooth nonparametric regression for two or more variables. *Canadian Journal of Statistics 34*, 4 (2006), 535–561. 4

[FG12] FORERO P. A., GIANNAKIS G. B.: Sparsity-exploiting robust multidimensional scaling. *IEEE Transactions on Signal Processing 60*, 8 (2012), 4118–4134. 2

[GJP*14] GENTON M. G., JOHNSON C., POTTER K., STENCHIKOV G., SUN Y.: Surface boxplots. *Stat 3*, 1 (2014), 1–11. 3

[HS10] HYNDMAN R. J., SHANG H. L.: Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics 19*, 1 (2010), 29–45. 3, 6, 7

[Hyn96] HYNDMAN R. J.: Computing and graphing highest density regions. *The American Statistician 50*, 2 (1996), 120–126. 3

[LC10] LECUN Y., CORTES C.: MNIST handwritten digit database. URL: http://yann.lecun.com/exdb/mnist/ [cited 2016-01-14 14:24:11]. 6

[Lic13] LICHMAN M.: UCI machine learning repository, 2013. URL: http://archive.ics.uci.edu/ml. 1, 7, 8

[LMW*17] LIU S., MALJOVEC D., WANG B., BREMER P.-T., PASCUCCI V.: Visualizing high-dimensional data: Advances in the past decade. *IEEE transactions on visualization and computer graphics 23*, 3 (2017), 1249–1268. 1

[LPR09] LÓPEZ-PINTADO S., ROMO J.: On the concept of depth for functional data. *Journal of the American Statistical Association 104*, 486 (2009), 718–734. 2, 3

[LPSLG14] LÓPEZ-PINTADO S., SUN Y., LIN J., GENTON M.: Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification* (2014), 1–18. 3

[McG66] MCGEE V. E.: The multidimensional analysis of 'elastic' distances. *British Journal of Mathematical and Statistical Psychology 19*, 2 (1966), 181–196. 3

[MLL12] MAHMOUD S. M., LOTFI A., LANGENSIEPEN C.: User activities outlier detection system using principal component analysis and fuzzy rule-based system. In *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments* (2012), ACM, p. 26. 1

[MWK14] MIRZARGAR M., WHITAKER R., KIRBY R.: Curve boxplot: Generalization of boxplot for ensembles of curves. *Visualization and Computer Graphics, IEEE Transactions on 20*, 12 (Dec 2014), 2654–2663. 2, 3

[MWK17] MIRZARGAR M., WHITAKER R. T., KIRBY R. M.: Exploration of heterogeneous data using robust similarity. *arXiv preprint arXiv:1710.02862* (2017). 8, 9

[RMR*17] RAJ M., MIRZARGAR M., RICCI R., KIRBY R. M., WHITAKER R. T.: Path boxplots: A method for characterizing uncertainty in path ensembles on a graph. *Journal of Computational and Graphical Statistics 26*, 2 (2017), 243–252. 2, 3

[RRT99] ROUSSEEUW P. J., RUTS I., TUKEY J. W.: The bagplot: a bivariate boxplot. *The American Statistician 53*, 4 (1999), 382–387. 1, 2, 3

[RW17] RAJ M., WHITAKER R. T.: Anisotropic radial layout for visualizing centrality and structure in graphs. *arXiv preprint arXiv:1709.00804* (2017). 2, 5

[SG11] SUN Y., GENTON M. G.: Functional boxplots. *Journal of Computational and Graphical Statistics 20*, 2 (2011), 316–334. 3, 5, 6

[SHS16] SHANG H. L., HYNDMAN R. J., SHANG M. H. L.: Package 'rainbow'. 2

[SL89] SPENCE I., LEWANDOWSKY S.: Robust multidimensional scaling. *Psychometrika 54*, 3 (1989), 501–513. 2

[SSM97] SCHÖLKOPF B., SMOLA A., MÜLLER K.-R.: Kernel principal component analysis. In *International Conference on Artificial Neural Networks* (1997), Springer, pp. 583–588. 1

[TdJD*13] TENNEKES M., DE JONGE E., DAAS P. J., ET AL.: Visualizing and inspecting large datasets with tableplots. *Journal of Data Science 11*, 1 (2013), 43–58. 10

[Tuk75] TUKEY J. W.: Mathematics and the picturing of data. 2, 3, 5, 6

[VSKB10] VISHWANATHAN S. V. N., SCHRAUDOLPH N. N., KONDOR R., BORGWARDT K. M.: Graph kernels. *Journal of Machine Learning Research 11*, Apr (2010), 1201–1242. 10

[Wil17] WILKINSON L.: Visualizing big data outliers through distributed aggregation. *IEEE transactions on visualization and computer graphics* (2017). 1, 9

[Win04] WINNER L.: UFO encounters, 2004. URL: http://www.stat.ufl.edu/~winner/datasets.html [cited 2017-11-29 10:55:11]. 8

[WMK13] WHITAKER R. T., MIRZARGAR M., KIRBY R. M.: Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec. 2013), 2713–2722. 2, 3, 5

[YNMX17] YANG V., NGUYEN H., MATLOFF N., XIE Y.: Top-frequency parallel coordinates plots. *CoRR abs/1709.00665* (2017). URL: http://arxiv.org/abs/1709.00665, arXiv:1709.00665. 10

[ZS88] ZWITTER M., SOKLIC M.: Breast cancer data, 1988. URL: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer. 8

[ZS00] ZUO Y., SERFLING R.: General notions of statistical depth function. *Ann. Statist 28* (2000), 461–482. 2