1-2019

Improved N-dimensional Data Visualization from Hyper-radial Values

Todd J. Paciencia

Trevor J. Bihl
Air Force Research Laboratory

Kenneth W. Bauer
Air Force Institute of Technology

Research Article

Journal of Algorithms & Computational Technology

Improved N-dimensional data visualization from hyper-radial values

Journal of Algorithms & Computational Technology
Volume 13: 1–20

DOI: 10.1177/1748302619873602
journals.sagepub.com/home/act
SAGE

Todd Paciencia[1], Trevor Bihl[2] and Kenneth Bauer[3]

Abstract

Higher-dimensional data, which is becoming common in many disciplines due to big data problems, are inherently difficult to visualize in a meaningful way. While many visualization methods exist, they are often difficult to interpret, involve multiple plots and overlaid points, or require simultaneous interpretations. This research adapts and extends hyper-radial visualization, a technique used to visualize Pareto fronts in multi-objective optimizations, to become an n-dimensional visualization tool. Hyper-radial visualization is seen to offer many advantages by presenting a low-dimensionality representation of data through easily understood calculations. First, hyper-radial visualization is extended for use with general multivariate data. Second, a method is developed by which to optimally determine groupings of the data for use in hyper-radial visualization to create a meaningful visualization based on class separation and geometric properties. Finally, this optimal visualization is expanded from two to three dimensions in order to support even higher-dimensional data. The utility of this work is illustrated by examples using seven datasets of varying sizes, ranging in dimensionality from Fisher Iris with 150 observations, 4 features, and 3 classes to the Mixed National Institute of Standards and Technology data with 60,000 observations, 717 non-zero features, and 10 classes.

Keywords

Data visualization, multivariate visualization, statistical graphics, visualization techniques, dimensionality reduction

Received 9 October 2017; Revised received 16 April 2019; accepted 16 June 2019

Introduction

High-dimensional data are naturally difficult to visualize in a meaningful way, as anything with more than four dimensions provides challenges.[1] Unfortunately, many real-world datasets have much greater than four dimensions and have complex interactions between features, making a simple plotting of feature subsets impractical for most purposes. While visual data mining can be used to find structures in datasets,[2] multivariate data complicates visualizations through the presence of those many features which have different interactions with other features.

Appropriate visualizations are frequently critical in data analysis, adding meaning, and displaying results, with best practices providing relatively simple and clear output to the audience.[3,4] Additionally, visualization can provide confidence in data exploration since visualizations are frequently more intuitive than complex algorithms.[5] For the purposes of this research, we are interested in being able to utilize an interpretable visualization in order to identify general characteristics of a multivariate dataset when little is known about its underlying structure. The extent of class overlap, discriminatory features, and presence of outliers and clusters in the data are all useful to visualize. In the application of classification, visualization may provide insight into class structure and the linearity of decision boundaries.

Various methods have therefore been proposed for visualizing multi-dimensional datasets. However, issues exist with these methods; some become computationally expensive as the number of data features increases,

[1]Headquarters U.S. Air Force, Washington, DC, USA
[2]Air Force Research Laboratory, Sensors Directorate, Wright-Patterson Air Force Base, OH, USA
[3]Air Force Institute of Technology, Wright-Patterson Air Force Base, OH, USA

Corresponding author:
Todd Paciencia, United States Air Force, 1570 Air Force Pentagon, Washington, DC 20330, USA.
Email: todd.j.paciencia.mil@mail.mil

while others are frequently not intuitive or do not lend themselves to the visualization of many data features. Surveys of various methods include those by Grinstein et al.,[6] Keim,[5] Kromesch and Juhasz,[7] Chan,[8] and Kehrer and Hauser.[9] Mühlbacher et al. presented a survey of frequently used algorithms and their fulfillment of certain visual analytic requirements.[10] The presented hyper-radial visualization (HRV) method is incidentally more user-friendly than many of these methods, e.g. neural networks, k-means, support vector machines, and t-Distributed Stochastic Neighborhood Embedding (t-SNE), in that HRV's basic operation and computations are straightforward, easy to implement, and if coded properly, can allow for some degree of interactivity as the visualization is built.

The HRV concept was originally proposed by Chiu and Bloebaum for visualization of Pareto frontiers in multi-objective optimization problems.[11] Herein, an efficient n-dimensional multivariate data visualization version of HRV is presented; this method is powerful in that data features are only aggregated, rather than transformed, to create the resulting visualization. Whereas HRV was originally designed for comparison of competing optimal designs, we broaden its use for visualizing class and exemplar characteristics in multivariate data. In order to improve the visualization, we also present optimization strategies to generate the groups required for both supervised and unsupervised cases. Now, as the number of features increases, any two-dimensional visualization becomes inherently limited in being able to display the information present. Here, the authors also create a three-dimensional version to enable visualization for larger numbers of features.

For this paper, example n-dimensional data are presented and existing visualization methods are reviewed, followed by these contributions (in order):

1. The HRV method is extended to multivariate data.
2. An optimal group algorithm is developed for the HRV visualization, both in the event of having and not having class information.
3. A three-dimensional version of HRV is developed incorporating the optimization strategies.

Example datasets

Seven example datasets, described in Table 1, are employed to illustrate, evaluate, and compare our HRV methods to existing visualization methods. These datasets range in size from 150 observations with 4 features and 3 classes in Fisher Iris,[12] to 60,000 observations, 717 (non-zero) features, and 10 classes in Mixed National Institute of Standards and Technology (MNIST).[13] All datasets have multiple classes, ranging from 2 to 10. Typically, data features correspond to measurements, e.g. Fisher Iris contains dimensional measurement of Iris flower petal and sepals.[12] Fisher Iris, in particular, is a common dataset used for visualization comparison.[6,14] In general, the datasets were taken "as is"; however, 16 missing values in the Breast Cancer dataset were imputed via L_1 nearest-neighbor approach within-class. Further details are necessary to understand the MNIST and Pavia datasets. MNIST contains data corresponding to visualizing hand-written numerals, and therefore all features are pixels in an image.[13] For data quality purposes, features (pixels) with zero range were removed. Pavia considers a 610×340-pixel hyperspectral image (HSI) from the ROSIS sensor, capturing bands between approximately 0.43 and 0.86 μm.[15] In HSI, each pixel of an image has an associated spectral signature over a set of bands, or discrete intervals on the electromagnetic spectrum.

These sets were chosen to showcase flexibility to number of exemplars, number of features, number of classes, and general data complexity. Since Fisher Iris[12] is both a commonly used dataset and among the smallest datasets examined herein, it will be presented first to show the disadvantages of other methods. Understanding the relative complexity and an ability to generalize is important. Although a direct numerical comparison of complexity for these datasets is difficult, an extended Fisher ratio, from Gu et al.,[19] for c classes

Table 1. Data under analysis.

Dataset	Number of classes	Number of features	Number of observations	Extended Fisher ratio (/p)
Insects[16,17]	4	3	36	2.08 (0.69)
Fisher[12]	3	4	150	30.78 (7.70)
Escherichia coli[18]	8	7	336	11.33 (1.62)
Breast Cancer (Diagnostic)[18]	2	9	699	10.93 (1.21)
Wine[18]	3	13	178	13.93 (1.07)
Pavia University (HSI) [15]	10	103	207,400	18.54 (0.17)
MNIST (Training)[13]	10	717	60,000	106.60 (0.15)

HSI: hyperspectral image; MNIST: Mixed National Institute of Standards and Technology.

summing the Fisher scores of each feature is provided. This is

$$\sum_{i=1}^{p}\left(\frac{\sum_{j=1}^{c} n_j(\mu_{ji}-\mu_{*i})^2}{\sum_{j=1}^{c} n_j\sigma_{ji}^2}\right) \quad (1)$$

where n_j is the number of exemplars in class j, σ_{ji}^2 is the sample variance of feature i in class j, μ_{ji} is the sample mean of feature i in class j, and μ_{*i} is the sample mean of feature i over all classes. As the visualizations later also imply, the Fisher dataset has the best general class structure among these datasets.

Existing visualizations

Various methods are in literature and practice for visualizing data, but all carry limitations. Though Lengler and Eppler created a periodic table of 100 visualizations to put some structure towards situational use, that structure does not provide a detailed explanation of the limitations of any given technique on a specific dataset.[20] Therefore, we discuss several techniques and their limitations here. Scatterplots are one common method, where each feature is plotted against another feature and simultaneous interpretations are posited.[7] Alternatively, features can be plotted three at-a-time to create three-dimensional figures. However, both of these methods can become difficult to interpret as the number of features or observations increase. For example, even with a relatively small number of features and observations, scatterplots are difficult to interpret, as seen in Figure 1(a) for Fisher Iris.

Parallel coordinates is another commonly used method, where features are normalized according to their range and then each exemplar is plotted as a line of its features.[7,21] To illustrate this method, Fisher Iris is visualized using parallel coordinates in Figure 1(b). It is apparent that even Fisher Iris is not easy to analyze and interpret with these coordinates due to many overlapping lines. Logically, more complex and larger datasets would be increasingly difficult to visualize with this tool, a problem described by Dang et al. as overplotting.[22] While variants of parallel coordinates also exist, such as connecting normalized feature values radiating from the center of a circle akin to a radar graph,[7] or using parallel dual plots,[23] these results can still be difficult to interpret due to overplotting.

Many additional visualization techniques exist. These include, in part, iconographic (or glyph) displays, multi-line graphs, by-feature heat maps, logic diagrams of features, survey plots of features, and hierarchal methods.[6,8] Additional methods include dimensional stacking,[1] multiple frames,[24] and nonlinear magnification.[24] Mosaic matrices,[25] using a hyperbox[26] and table lens,[27] are also used. RadViz places dimensional anchors (the features) around a circle, with spring constants utilized to represent relational values among points.[14] PolyViz is a similar construct, with each feature anchored instead as a line.[6] This is depicted for Fisher Iris in Figure 2. All of the methods mentioned have obvious interpretation, overplotting, and clutter issues as the number of features and/or exemplars grows.

Some visualization techniques have been developed in the field of multi-objective optimization to be able to compare Pareto optimal solutions for problems with more than three objectives. There, these objective functions are optimized simultaneously. In order to determine optimality, optimal trade-offs are maintained, where a solution is Pareto optimal if no other feasible point is better in all objectives. Hyperspace Diagonal Counting is a method based on the premise of Cantor's counting method, mapping exemplars to a line by counting along hyperdiagonal bins that move away from the origin.[28] However, this method becomes inefficient to compute as the number of features and exemplars grow, and may gravitate values toward the axes thus limiting its usefulness.[29] Another technique, which we leverage herein, is presented in the next section.

Dimensionality reduction is another class of techniques that can be used within visualization methods to try and reduce the amount of information via either feature extraction or feature selection. Of note here are feature extraction methods that transform data to a different space. Principal component analysis (PCA), for instance, generates projection vectors that account for variability found in the data.[30] Thus, data can be projected into a smaller number of dimensions (new features) while retaining a percentage of the total variance. Unfortunately, PCA projections do not guarantee that characteristics of the data, such as distances between points, are maintained. Instead, the Johnson–Lindenstrauss theorem shows that for any $0 < \epsilon < 1$, any set of n points X in $\mathbb{R}^p$, and $p \geq k \geq 4\left(\epsilon^2/2 - \epsilon^3/3\right)^{-1} \ln n$, there exists a map $f: \mathbb{R}^p \rightarrow \mathbb{R}^k$ that can be found in randomized polynomial time such that for all $u, v \in X \subset \mathbb{R}^p$, $(1-\epsilon)||u-v||^2 \leq ||f(u)-f(v)||^2 \leq (1+\epsilon)||u-v||^2$.[31,32] This theorem implies the existence of a mapping that could be found that would maintain the distances between points in the mapped space and the original space. Such a mapping would be very powerful for the purposes of visualization. Unfortunately, a way to explicitly generate this mapping has yet to be determined, and the required bound on k relative to a small ϵ can still require large dimensionality. Achlioptas took it a step further and determined a projection matrix that

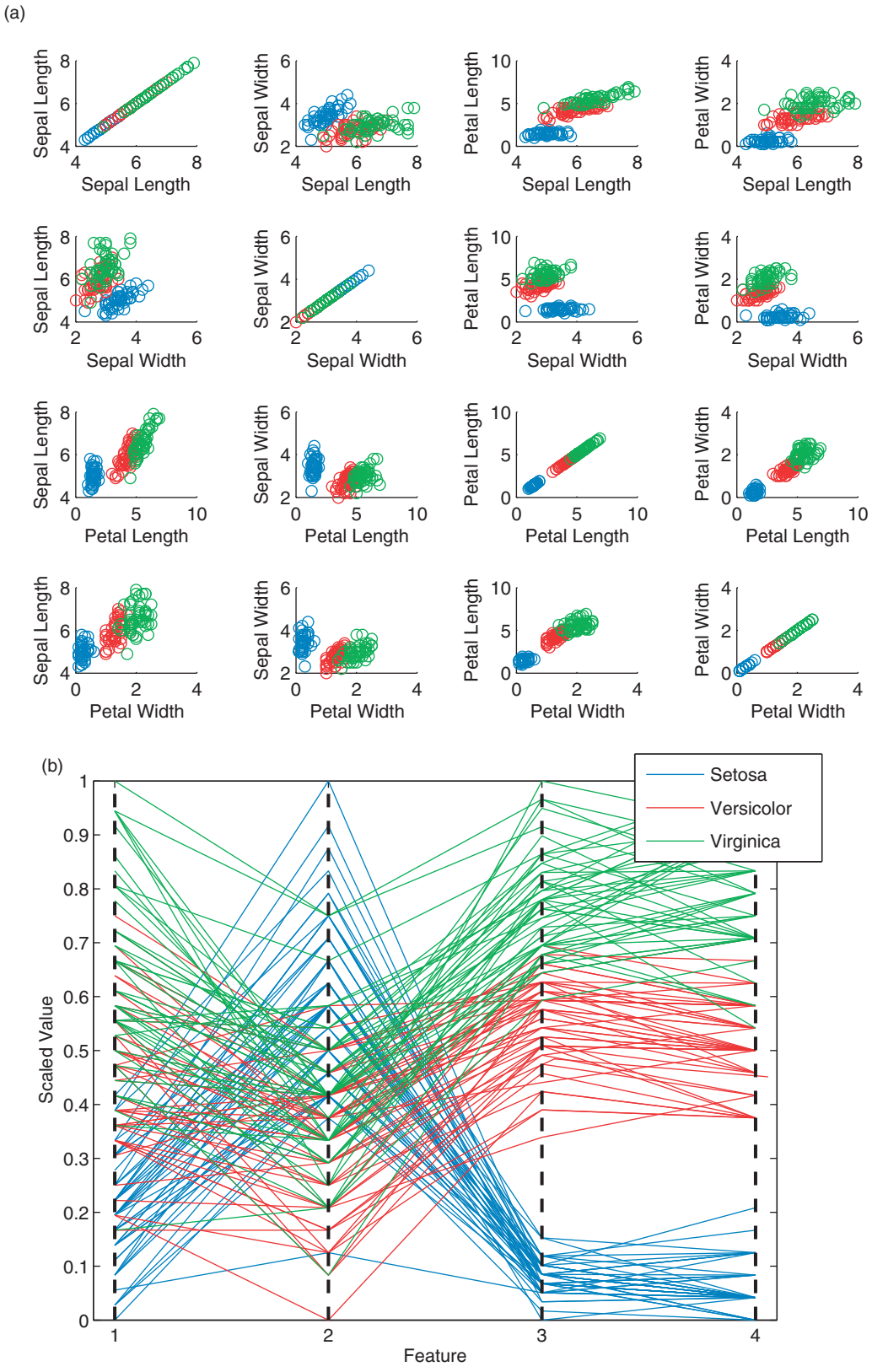

Figure 1. (a) Fisher Iris feature-by-feature scatterplots and (b) parallel coordinate representation.

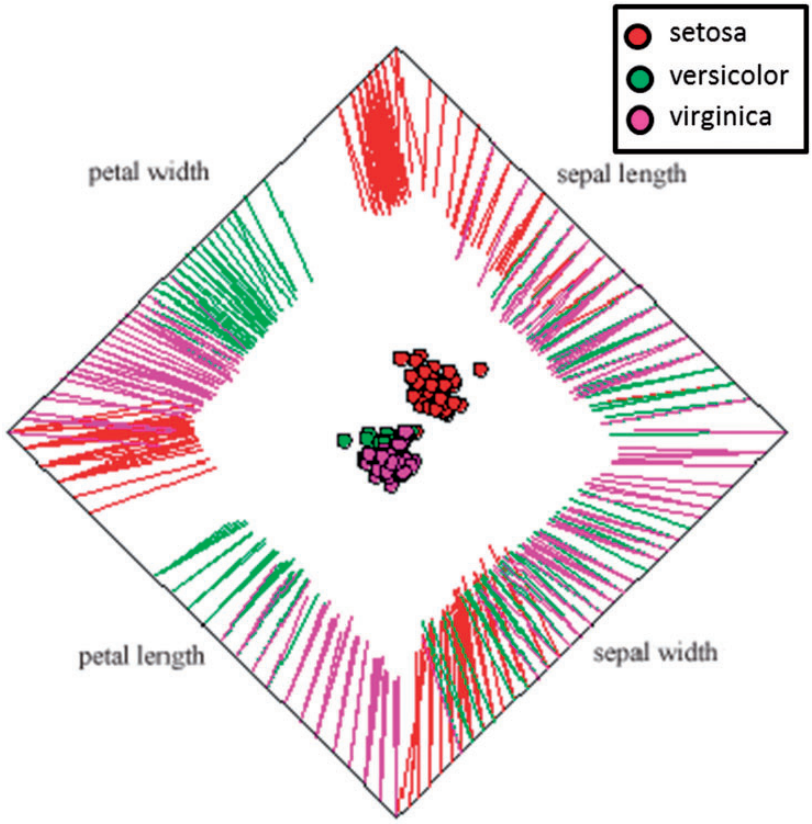

Figure 2. Fisher Iris PolyViz visualization.
Source: Reprinted from ACM Press.[6]

satisfies the Johnson–Lindenstrauss theorem with a modified bound on k

$$k \geq \left(4 - \frac{2\ln\ (1-q)}{\ln\ n}\right)\left(\epsilon^2/2 - \epsilon^3/3\right)^{-1}\ln\ n \qquad (2)$$

with probability at least q.[33] This is still problematic, in that the lower bound for k, k_0, is still high even for large ϵ and low q. For example, if $n = 200$, $\epsilon = 0.9$, and $q = 0.1$, $k \geq 132$. In fact, we can note from Figure 3, a plot of this bound over worst-case ranges of n, ϵ, and q that the bound increases as n and q increase, and as ϵ decreases. Multi-dimensional scaling is another technique that tries to approximate the Johnson–Lindenstrauss mapping, but with Euclidean distance as the similarity metric, the embedding is the same as PCA scores and does not guarantee maintenance of distances.[34] Numerous other dimension reduction techniques exist, but they typically involve using weighted combinations or transformations of the features, making them very difficult to interpret in terms of how they relate to the original data.

Bertini et al. suggested the evaluation of high-dimensional data visualizations via (1) the extent to which data groupings are maintained, (2) the extent to which systematic changes in one dimension are accompanied by changes in others, (3) the maintenance of outliers, (4) the level of clutter or crowding that could make interpretation difficult, (5) the extent to which feature information is preserved, and (6) any remaining aspects that may add complexity to the visualization.[35] Despite the abundance of visualization techniques that exist in the literature, the authors suggest that few, if any, meet a level of quality for several of these metrics. Therefore, we seek to leverage an existing visualization from the optimization field in order to attempt to satisfy, at a minimum, maintenance of data groupings and outliers, reduced clutter relative to classes, interpretability relative to original features, and minimal complexity of the visualization.

Hyper-radial visualization

Let F_i denote the ith feature (column) of the $N \times p$ exemplar-by-feature dataset X. In order to create a hyper-radial method for general data similar to the work of Chiu and Bloebaum in multi-objective

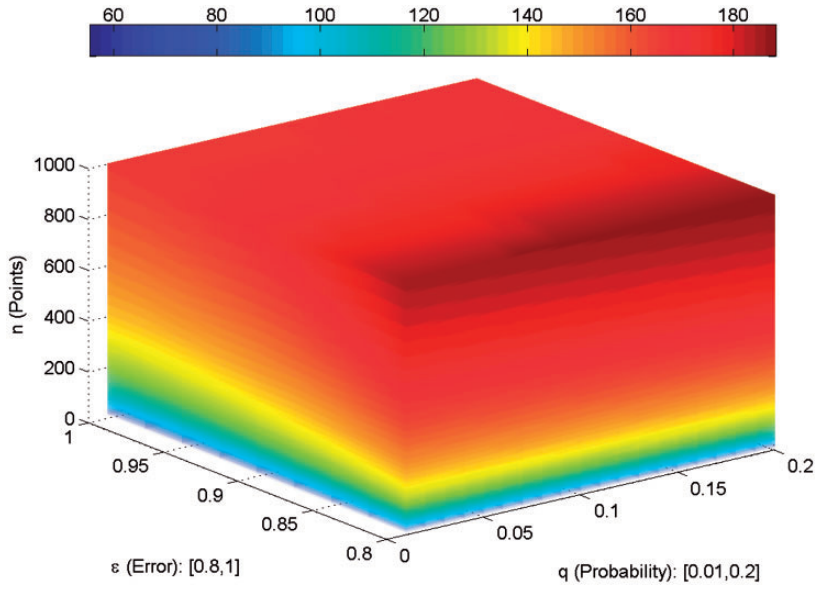

Figure 3. k_0 values as a function of error ϵ, probability q, and n observations in reference to equation (2) and the projection from Achlioptas.[33]

optimization,[11] first, we normalize each feature according to

$$\tilde{F}_i = \frac{F_i - F_{i,min}}{F_{i,max} - F_{i,min}} \in [0, 1] \tag{3}$$

for $i = 1, \ldots, p$, where $F_{i,min}$ and $F_{i,max}$ are the minimum and maximum values of the exemplars in that feature. Although this changes the scale of features relative to one another, it maintains the information found within the feature and also later ensures values in the interval $[0, 1]$ for the visualization, preventing outliers from skewing the visualization too greatly.

Next, features are grouped into two sets, most simply

$$G_1 = \left\{\tilde{F}_1, \tilde{F}_2, \ldots, \tilde{F}_s\right\} \text{ and } G_2 = \left\{\tilde{F}_{s+1}, \tilde{F}_{s+2}, \ldots, \tilde{F}_p\right\} \tag{4}$$

where $s = \lceil p/2 \rceil$. In their work with objective function data, Chiu and Bloebaum did not choose such groups in any special way.[11] For each group, a hyper-radial calculation (HRC) value is computed for each exemplar as

$$HRC_j = \sqrt{\frac{\sum_{i \in G_j} \tilde{F}_i^2}{n_j}} \tag{5}$$

where $j = 1$ or 2 for G_j, and n_j is the number of features in group j. To maintain an unbiased representation, the two groups are kept equal in size. Thus, for an odd number of features, one group is given a dummy zero objective. With two groups, points can be plotted two-dimensionally using the *HRC* values. Finally, curves of constant distance from the origin (minimum feature values) are added to the plot.

Fisher Iris is visualized through HRV in Figure 4(a). As can be seen, this visualization already clearly depicts some class boundaries. The axes are annotated with the grouping number, e.g. G_1, and the features grouped on a given axis, e.g. $F : 1\,3$ for feature 1 and feature 3. Additionally, the data are not plotted through a myriad of plots or overlapping lines, as shown in Figure 1.

This technique is powerful in that it is easily interpretable and calculable. In reality, the *HRC* values are just a weighted Euclidean distance, or hyper-radial, of the groups of normalized features from their minimum values. This is easier to directly interpret than PCA, e.g. where each axis is a different weighted sum of features. With HRV, the geometry of the data is essentially maintained through a polar plotting approach without any true transformation of the data. Similarity between exemplars is maintained for each feature group within a scaled factor. Minimums occur at 0, and maximums at 1, making it easy to relate positioning of solutions to one another.

Improving HRV

However, there are also limitations to this initial HRV methodology when used as a data visualization tool. The groups aggregate information from the features,

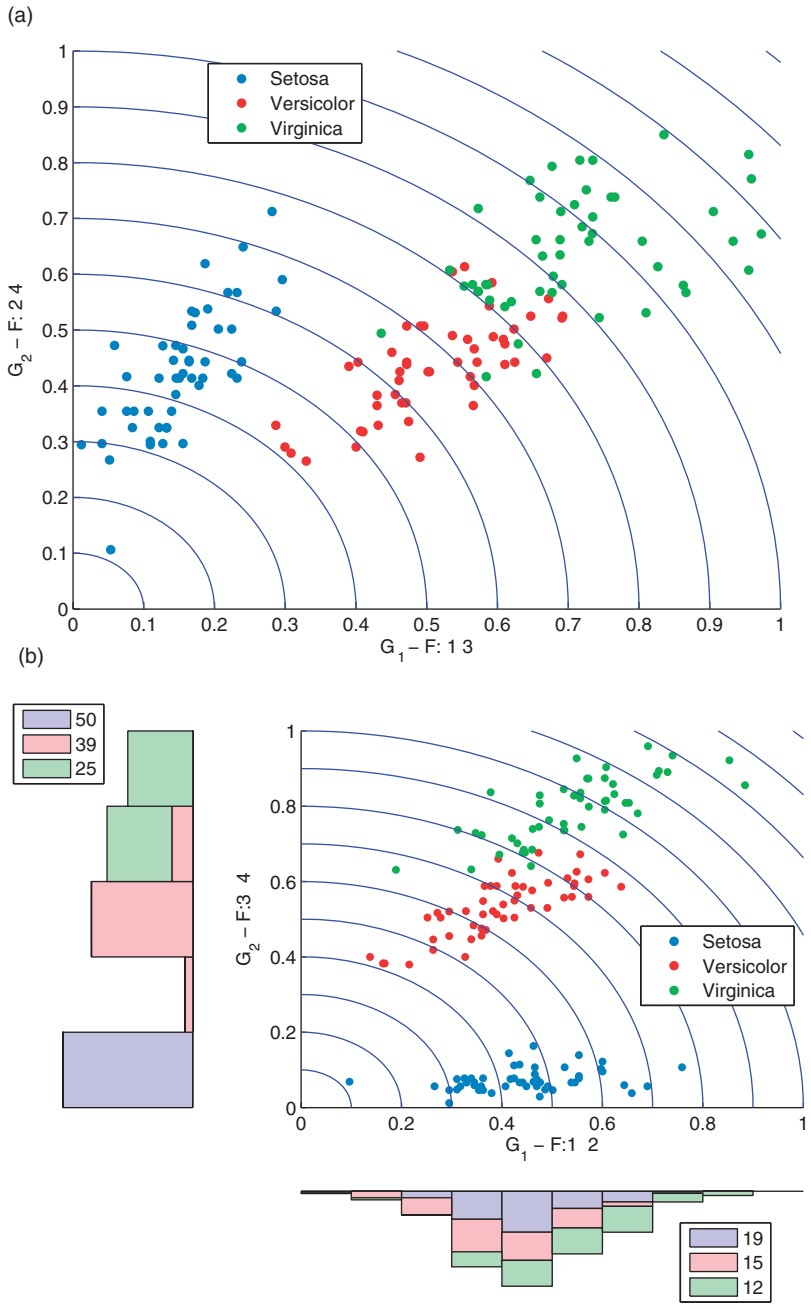

Figure 4. (a) A Fisher Iris HRV representation and (b) instead with "optimal" groups.
HRV: hyper-radial visualization.

and so different data points can possibly map to the same point in the visualization. Further, the membership of each group can have a significant impact on the visualization. As overall visualization crowding is difficult to avoid in a low number of dimensions, we are more concerned with class and outlier characteristics of the data, and the issue of group membership.

We propose using the adapted HRV method with multivariate data, after the addition of a few further modifications. Adding stacked-bar histograms to each *HRC* axis in the visualization can serve as an additional way to detect and see information as no matter the visualization, any overlap of high-volume, high-dimensional data can be difficult to determine in two or three dimensions. In the following sections, we will also introduce a method to choose optimal groupings and a third HRC axis for use with larger-dimensional data.

Figure 4(b) displays an HRV again for Fisher Iris with stacked-bar histograms and different groups from Figure 4(a), where the histogram legends denote the

largest number of class exemplars in any bin, for a relative size comparison. It is clear from the visualization that the largest separation can be achieved using the third and fourth features (petal length and width). The class boundaries are also now obvious, and this demonstrates that HRV presents an effective means to visualize this four-dimensional data in two dimensions.

As alluded to, the success of the visualization and these improvements to HRV, for our purposes, are still entirely dependent on a proper choice of grouped features for the *HRC* computations. Therefore, next, we discuss strategies with which to find optimal, useful groupings.

Determining optimal groupings

Given some objective J_t, here we use t simply as an index with which to reference a specific objective function, finding an optimal grouping in two dimensions can be formulated as shown in equation (6), where x_i is the value of the exemplar x in the ith feature, $\tilde{x}_i$ is its normalized value, and for later notation simplicity, we use v_j to denote the jth HRC axis coordinates HRC_j. If p is odd, recall that a dummy zero feature is added to one of the groups to keep the visualization unbiased. This formulation is robust to that adjustment. Additionally, this new binary set formulation enables us to develop strategies to find optimal groups. However, as the objective functions we use and develop here are highly non-linear directly or as a result of also having the non-linear HRC values input, and because group selection is binary, this problem is not trivial. That is, linear under-estimators[36] or pseudo-Boolean methods[37] cannot necessarily be used here to simplify or speed the non-linear optimization. Instead, we develop a simple heuristic method to be able to efficiently generate an optimal, or pseudo-optimal, visualization for data when complete enumeration is not an option. In Appendix 1, we compare this method to optimizing a relaxed version of the problem using non-linear programming methods, relaxing the binary constraints until a final group selection. First, we will present objectives to use for J_t

$$\begin{aligned}
&\max\ J_t(X,y) = J_t(v_1(X), v_2(X)) \\
&\text{subject to } \sum_{i=1}^{p} y_i = \lceil p/2 \rceil \\
&y_i \in \{0,1\}, \quad \text{for } i = 1,\ldots,p \\
&x \in X, \quad \text{where,} \\
&v_1(x) = HRC_1(x) = \sqrt{\frac{\sum_{i=1}^{p} y_i \tilde{x}_i^2}{\sum_{i=1}^{p} y_i}}, \text{ and} \\
&v_2(x) = HRC_2(x) = \sqrt{\frac{\sum_{i=1}^{p} (1-y_i) \tilde{x}_i^2}{\sum_{i=1}^{p} (1-y_i)}}
\end{aligned} \tag{6}$$

Supervised training. In the case where class information is known, we propose that the Rayleigh coefficient from multiple Fisher discriminant analysis (MDA) can be used as motivation to find groups with near-optimal class separation (or optimal in the linear sense). This coefficient is a ratio such that its maximization increases separation between class means and decreases the separation within class data.

In MDA, a set of $\min(c-1, p)$ optimal linear projection directions are desired, where c is the number of classes, in order to best separate the means of the projected classes and minimize their within-class variances. Linear directions are used in MDA because the equivalent non-linear problem would increase the size of the data due to the use of kernels in Kernel MDA, where this latter non-linear form also necessitates a good choice of kernel.[38–40] The within-class variance matrix of the visualization data is

$$S_W = \sum_{i=1}^{c} \sum_{x \in X_i} (v(x) - \mu_i)(v(x) - \mu_i)^T \tag{7}$$

where the subscripts denote the class, the HRC coordinates are in column vector form, and μ_i is the mean of the HRC coordinates, for exemplars in class i. The between-class variance matrix S_B is defined so that the total scatter in the visualization data is $S_B + S_W$. This defines

$$S_B = \sum_{i=1}^{c} n_i (\mu_i - \mu)(\mu_i - \mu)^T \tag{8}$$

where μ is the overall mean of the HRC coordinates and n_i reflects the number of exemplars in class i.[41] In MDA, data (here our visualization data) would be projected onto the multiple linear directions W, such that the following ratio would be maximal

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|} \tag{9}$$

This criterion, often noted as the Rayleigh coefficient or quotient,[39] is the equivalent ratio of between-class and within-class scatter for the projected space. Here, because $|A| = det(A) = \prod_l \lambda_l$, where λ_l are the eigenvalues of the matrix A, this ratio uses the products of the "variances" in the principal directions, or square of the hyperellipsoidal scattering volume.[41] Thus, a maximization serves to maximize the between-class scatter and minimize the within-class scatter in the projected space. From this criterion, the optimal projections can be found via an eigen-problem.

In our case, we do not wish to optimize to find optimal projection vectors. Instead, we wish to stay in our original visualization coordinates $v(x)$ so that the results are more interpretable, although the visualization can be applied to projections as well. That is, the axes are more easily understood if they only represent hyper-radials, rather than some other non-linear transformation, differently weighted projections, or a combination thereof. Thus, we use a form of equation (9) directly in our *HRC* space to form an optimization with similar intent for input into the formulation from equation (6). That is, we can use

$$J_1(v_1(X), v_2(X)) = \frac{|S_B|}{|S_W|} \tag{10}$$

where S_B and S_W are computed using the *HRC* coordinates for the input data exemplars. This maximizes the visualization between-class scatter and minimizes its within-class scatter in its principal directions. Unfortunately, this also means that we must solve our formulation rather than simply solving the eigenproblem for an optimal as in MDA.

J_1 may best linearly separate the data, but the resulting coordinates do not simultaneously seek to spread the data well across the axes and the determinants may yield very small values in the normalized space. For these reasons, we can optimize

$$J_2(v_1(X), v_2(X)) = \frac{tr(S_B)}{tr(S_W)} \tag{11}$$

where $tr(A) = \sum_i A_{ii} = \sum_l \lambda_l$. This formulation therefore still works to separate class means and minimize within-class scatter, but does so at an aggregation across the *HRC* axes because it uses a sum of the diagonal elements of S_B and S_W (and equivalently, the sum of eigenvalues).

With two groups and p features, there are $\binom{p}{\lceil p/2 \rceil}$ ways to assign features into groups. Therefore, for smaller p, it is possible to do complete enumeration to find optima, while for larger p, an optimization algorithm must be used. The advantage of using this optimization and objective functions is shown for Fisher Iris in Figure 4.

Unsupervised training. In the case where there is no class information, J_1 and J_2 are no longer useful unless predictions are made. Therefore, we propose a collection of objectives designed to spread the data maximally across the *HRC* space in various ways, under the assumption that doing so will help to reveal classes, overlaps, outliers, or other useful information.

The first technique is to maximize entropy over the *HRC* space, where maximal entropy indicates an even spread of data across the *HRC* dimensions. To do this, we define a grid of N_g centers over the $[0,1] \times [0,1]$ *HRC* axes. For each grid center point u, a density d_u is computed using radial basis functions as

$$d_u = \sum_{x \in X} \frac{1}{\sigma\sqrt{2\pi}} e^{||v(x)-u||^2/(2\sigma^2)} \tag{12}$$

where σ is a spread parameter that defines the decay of the basis. A grid center point with many nearby *HRC* solutions will have a higher density value. The densities are then converted to act like probabilities through normalization

$$\tilde{d}_u = \frac{d_u}{\sum_{u \in Grid} d_u} \tag{13}$$

The resulting entropy is defined as

$$H = -\sum_{u \in Grid} \tilde{d}_u \ln\ \tilde{d}_u \tag{14}$$

where $\ln\ N_g$ is the maximum possible value of H on the grid. Therefore, H can be scaled by this maximum to form an objective to maximize

$$J_3(v_1(X), v_2(X)) = \frac{H}{\ln\ N_g} \tag{15}$$

As with any Gaussian method, the value for σ can have a major effect. Fortunately, we know that the coordinates will always be $[0,1]$, enabling the choice for σ to be a desired sensitivity where issues may arise only if there are many outliers or singleton points in grid cells. Figure 5(a) shows the optimal for J_3 on the breast cancer data, using a 33×33 grid and $\sigma = 0.025$, along with the grid entropy density, in Figure 5(b). The benign and malignant classes are largely separable with a minor level of overlap, and the plot of the entropy densities shows clear evidence of these two classes, where a large number of benign exemplars are located near the lower bound of the second group. Minimizing J_3, instead of maximizing it, could also serve to aid outlier identification because the data would be made as Gaussian as possible.

Currently, the common σ over the grid provides a sensitivity threshold for the entropy surface. Alternatively, adaptive radial basis functions or adaptive kernel density estimation could be used to provide a more flexible entropy surface by allowing the spread parameter to

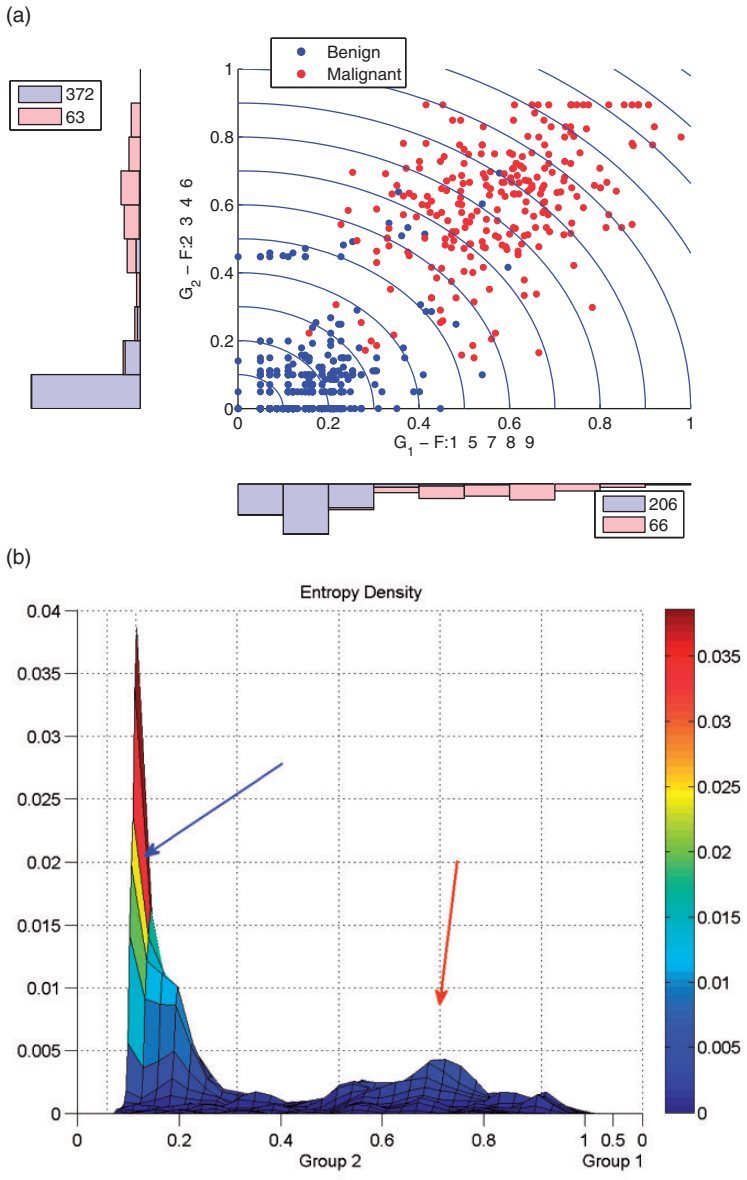

Figure 5. Breast cancer dataset: (a) visualization with optimal J_3 and (b) corresponding densities.

vary locally.[42] However, this method more precisely fits and smooths the surface to areas of high and low density, which may decrease or increase sensitivity in an unwanted manner. This also increases the computational effort needed to generate the entropy. Full evaluation of the effect of using such a method is of interest for future research.

There are a few other objectives that serve to spread the data points as best as possible in the *HRC* space, while specifically being useful to identify outliers. Maximizing the absolute value of the correlation between *HRC* coordinates best spreads the data along the (v_1, v_2) diagonal, but could also make the visualization very linear and make it harder to see characteristics of the data. However, a similar idea is to maximize the combined (v_1, v_2) spread. One such technique is to multiply the variances found in each direction

$$J_4(v_1(X), v_2(X)) = \prod_{i=1}^{2} Var(v_i(X)) \tag{16}$$

Another is to force the spread to the extremes in both directions simultaneously while avoiding bias in any one direction

$$J_5(v_1(X), v_2(X)) = \prod_{i=1}^{2} \left(\max_{x \in X} v_i(x) - \min_{x \in X} v_i(x) \right) \tag{17}$$

Yet another related objective is to minimize the absolute value of the correlation between coordinates.

These objectives may become limited when the outliers of interest are outliers only with respect to a very small number of features relative to the size of the groups.

As an example of a large-scale application, Figure 6 depicts the image truth and a sample of 100 signatures from each class for the Pavia University HSI. Figure 7(a) shows J_4 for this data. In both cases, a qualitative ColorBrewer[43,44] color scheme is used to provide distinction between classes. The groups found distinguish the non-background classes very well, and particularly of interest is how the painted metal sheets pixels are revealed to be significantly different. The visualization suggests that the background class contains material or mixes that are similar to some of the other classes. Several outliers are also obvious here, and the groups indicate a subset of features that make these outliers different. When compared to the primary components of PCA, shown in Figure 7(b), the visualization provides more separation between the classes and pixels, all while the axes are more directly interpretable.

Semi-supervised training. Additionally, HRV is suitable for semi-supervised analysis when class information is missing for some of the data, or if surrogate membership information is available. Two situations will be examined to illustrate: first, values computed from expected groupings, e.g. clustering algorithms, are considered; then, unknown or new data points are considered in the presence of some class information thereby using HRV to suggest possible class identities. HRV naturally lends itself to the first purpose, due to both HRV and methods such as k-means clustering being Euclidean distance-based. The second purpose involves using the supervised HRV methods with unknowns as an additional class and then visually determining the hypothetical class identities of the unknown observations.

For the clustering semi-supervised approach, the *Escherichia coli* dataset will be considered. While this dataset has known classes, for illustrative purposes, k-means clustering will be used to find suggested groups. *E. coli* consists of data from 7 features for 336 protein sequences and 8 classes (cellular component where each protein is found) collected on

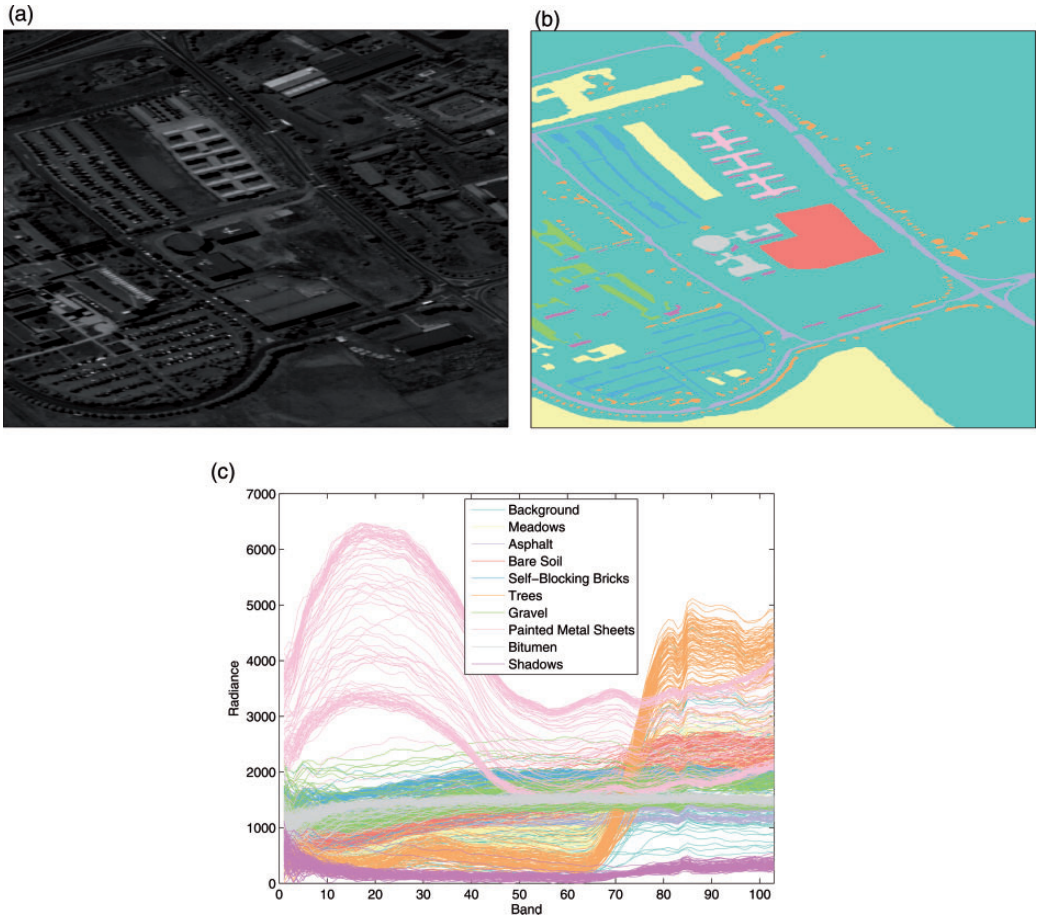

Figure 6. Pavia University HSI: (a) gray-scale image, (b) class truth, and (c) class spectral signatures samples.
HSI: hyperspectral image.

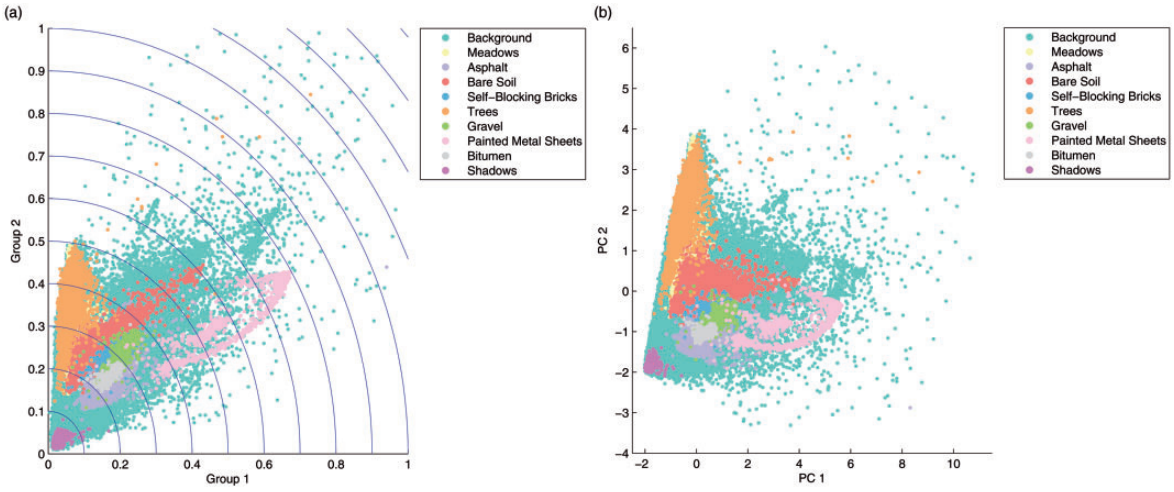

Figure 7. Pavia University: (a) J_4 HRV solution and (b) largest variance principal components.
HRV: hyper-radial visualization.

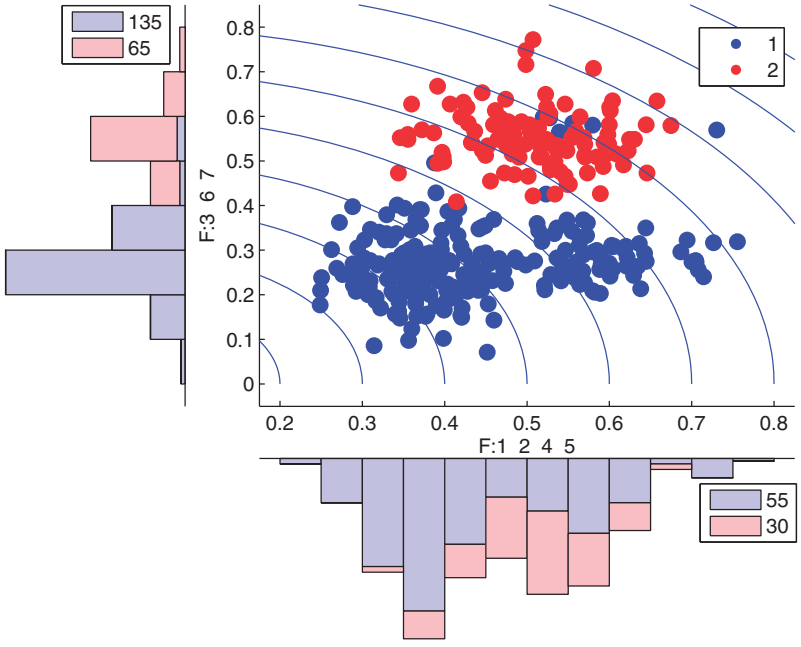

Figure 8. HRV applied to *Escherichia coli* $k = 2$, labeled with clustering result.
HRV: hyper-radial visualization.

Table 2. *Escherichia coli* class memberships in *k*-means clusters.

Protein localization site (class)	Cluster 1	Cluster 2
cp (cytoplasm)	99.70%	0.30%
im (inner membrane without signal sequence)	2.98%	97.02%
pp (periplasm)	99.40%	0.60%
imU (inner membrane, uncleavable signal sequence)	0.30%	99.70%
om (outer membrane)	99.70%	0.30%
omL (outer membrane lipoprotein)	100.00%	0.00%
imL (inner membrane lipoprotein)	50.00%	50.00%
imS (inner membrane, cleavable signal sequence)	50.00%	50.00%

various *E. coli* proteins.[45] Since eight classes can provide an over-abundance of visual information for interpretation, and because many of the classes have few exemplars, finding statistical groups in data through *k*-means may be justified. With $k = 2$ in *k*-means, two clusters are found with 229 exemplars in cluster 1 and 107 exemplars in cluster 2. These resulting clusters are depicted in Figure 8 using HRV, where a largely clean separation between clusters is evident. When comparing these clusters with the known *E. coli* classes, the groupings appear to also have logical sense with the membership information found in Table 2 describing the groups. Analyzing the class memberships in Table 2 indicates that the clusters fall along protein types with inner membrane localizations largely grouped together and outer membranes, periplasm, and cytoplasm grouped together. HRV provides further levels of detail, both in regard to class similarity and discriminatory features (Group 2 axis), that the cluster identities alone do not provide.

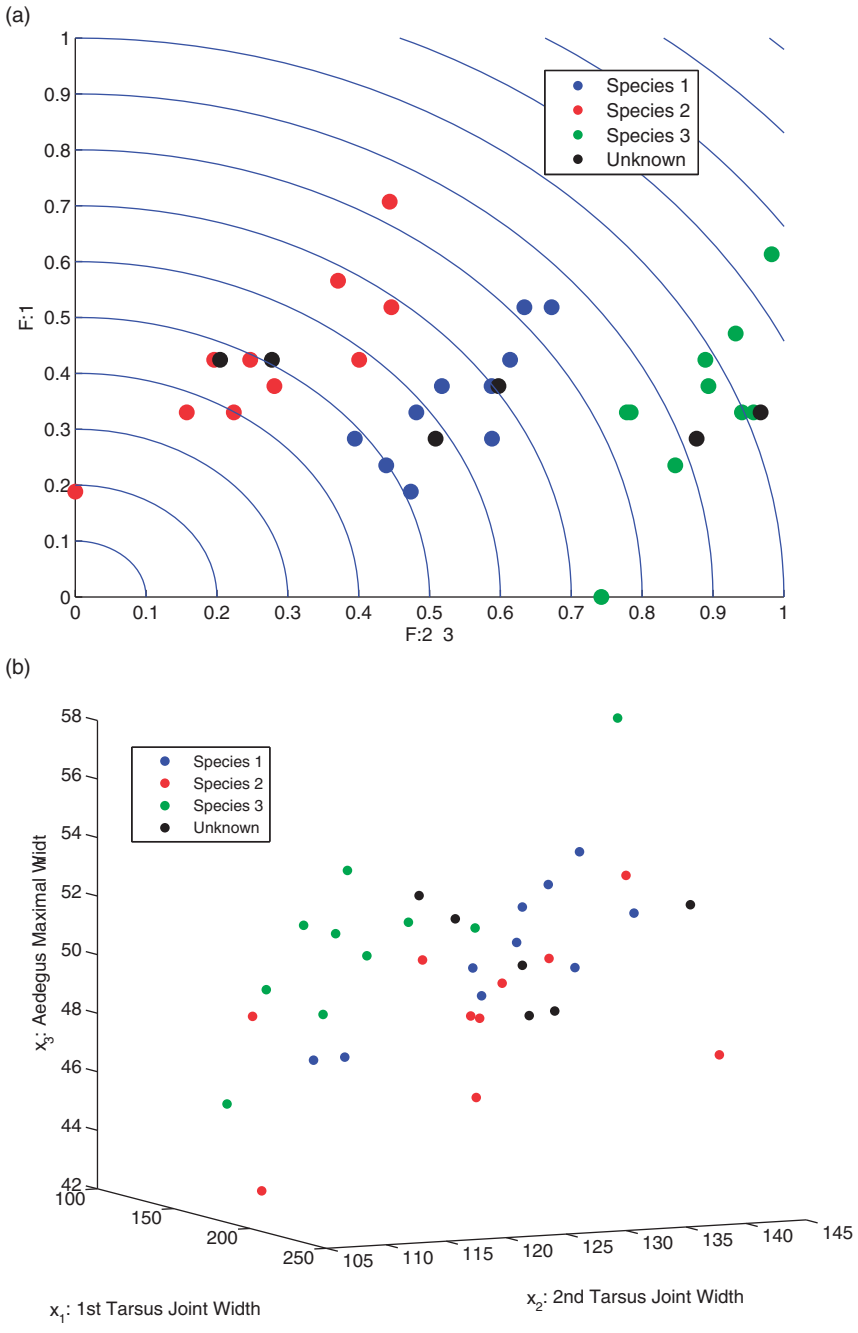

Figure 9. (a) HRV applied to insect data and (b) insect data feature scatterplot.
HRV: hyper-radial visualization.

The second type of semi-supervised approach with HRV involves the presence of unknown or new observations, and then using HRV to ascertain information on these unlabeled observations. The Insect dataset contains 30 observations of known species with six additional observations of unknown species.[16,17] The additional six observations are, however, known to belong to one of the known species. Lindsey et al. illustrated the utility of a multidimensional projection method to cluster the unknown species into the known groups.[16] However, using HRV can achieve the same performance without their intensive method as exemplified in Figure 9(a), where the insect data separate cleanly by class and the class assignment of unknowns appears logically distributed. Admittedly, the insect data are only three-dimensional, but the ability to identify the classes depends entirely on the rotation of any three-dimensional plot, as shown in Figure 9(b).

Three-dimensional HRV

As the number of classes, features, and exemplars increase, it becomes more challenging to display data in a meaningful way without transforming or projecting it, simply due to the amount of information that is being constrained to two dimensions. Unfortunately, a transformation or projection is not always intuitive to the intended audience for visualization, or may not be efficient to compute. As one example, van der Maaten and Hinton created a relatively successful cluster visualization of a 6000 exemplar subset of the MNIST dataset using their t-SNE algorithm.[46] However, t-SNE models Kullback–Liebler divergence between neighborhood conditional probabilities for all exemplars in the original and transformed spaces. Such an approach is computationally expensive, as the conditionals are computed for all exemplars and the transformed space is updated iteratively via a gradient approach. Further, feature information is lost and only a measure of aggregate proximity is maintained. The algorithm attempts to mitigate crowding of points, thus artificially adjusting the closeness of certain exemplars and clusters in the visualization.

In terms of HRV, we propose to help mitigate these issues by adding a third group for the *HRC* set of coordinates. All of the formulations and any heuristics easily adapt to incorporating the third group by adding another set of binary variables, and dummy features are used

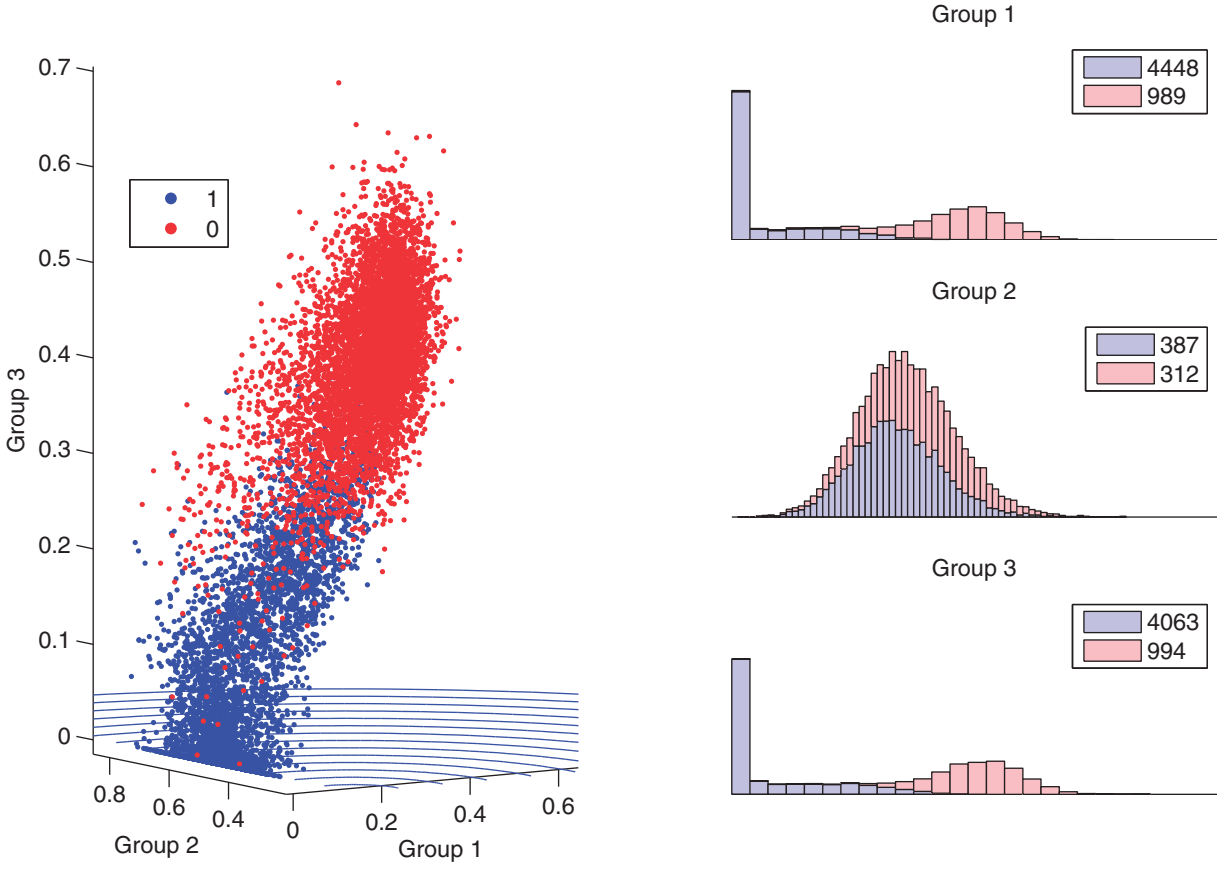

Figure 10. MNIST HRV using J_2 and the 0/1 classes.
HRV: hyper-radial visualization; MNIST: Mixed National Institute of Standards and Technology.

as needed to ensure equal group size. In order to expand the formulation to three groups, the binary constraints become those shown in equation (18)

$$\begin{aligned} &\sum_{i=1}^{p} y_i = \lceil p/3 \rceil \\ &\sum_{i=1}^{p} z_i = \lceil p/3 \rceil \\ &y_i + z_i \leq 1 \quad \text{for } i = 1, \ldots, p \\ &y_i, z_i \in \{0, 1\}, \quad \text{for } i = 1, \ldots, p \end{aligned} \tag{18}$$

As mentioned previously, as the number of classes, exemplars, and features grows, any visualization that tries to avoid true transformation will encounter issues due to the amount of information being constrained to an interpretable space. However, HRV can still be a useful tool. For example, consider the full training 0 and 1 classes from MNIST. Removing pixels that are 0 for all exemplars from both classes, the number of possible groupings is still $\binom{617}{206\,206\,205} = 617!/\left((206!)^2 205!\right)$. Figure 10 shows the visualization found for J_2 using only 8500 function evaluations of

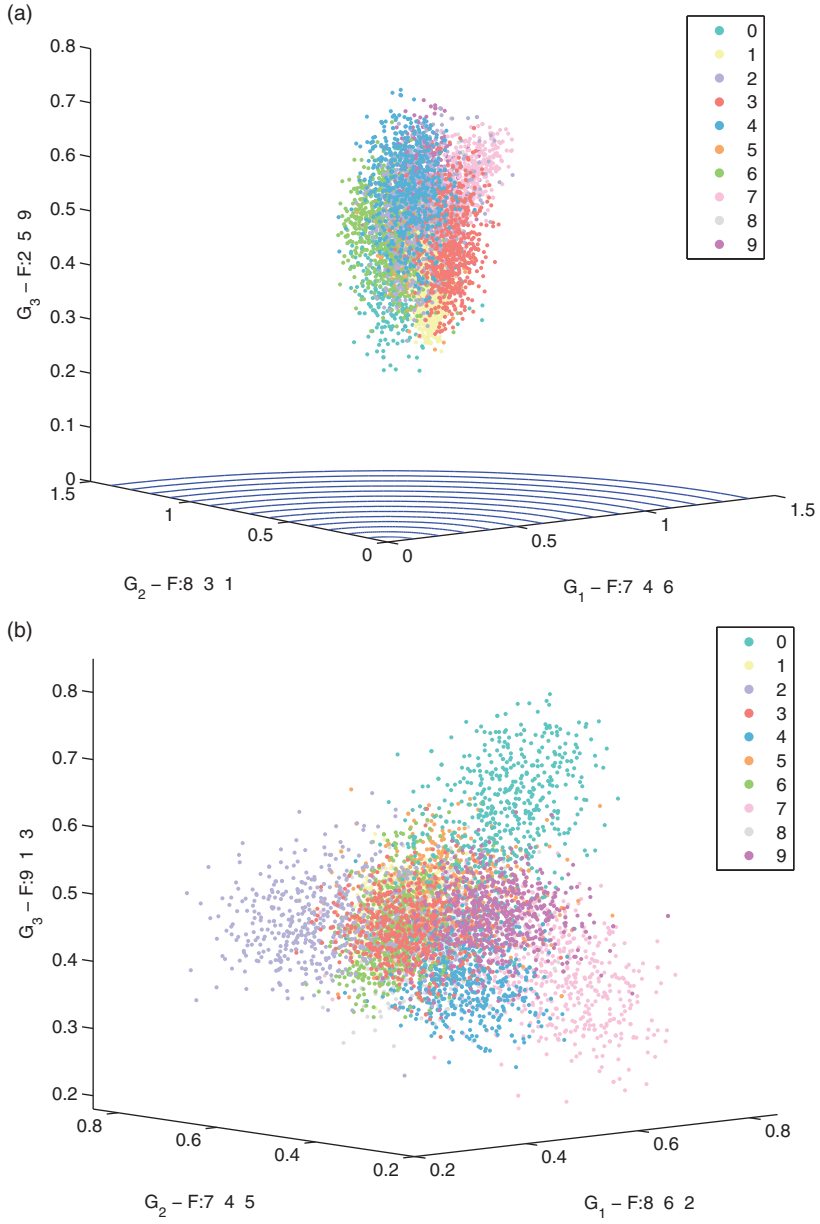

Figure 11. MNIST HRV representations using J_1: (a) principal component scores and (b) MDA scores.
HRV: hyper-radial visualization; MDA: multiple Fisher discriminant analysis; MNIST: Mixed National Institute of Standards and Technology.

a simple local search and random poll optimization algorithm, described fully in Appendix 1. Group 1 is highly discriminated and the 1-digits also present distinctly lower in Group 3. Furthermore, it is clear that there are two classes in the data from the histograms. Interestingly, in the unsupervised case, using J_5 often yielded the same visualization and two highly discriminated groups.

All of the desirable properties of HRV extend to the three-dimensional representation as each axis is still Euclidean-based for that group. In fact, using a third axis enables more distinction between points in the axis histograms. The only disadvantage to using three axes vice two is that the number of possible groupings is larger, and grows at a faster rate as a function of p, making the optimization potentially more difficult.

HRV can also be used to compare data projections. Here, we use a random sample of 600 exemplars from each class in MNIST. Figure 11 shows the J_1 optima on the principal component and MDA scores for the nine major components in each case. Whereas the PC scores are more compact and have significant overlap of some classes in any direction, the MDA scores break out better by class and are more spread. This better geometry from the MDA result might suggest the presence of multiple classes in an unsupervised setting. The better visualization from MDA would be expected to some level, as MDA provides the optimal linear projections for class separation. Additional comparisons to dimension reduction methods, and further investigation into the embedded optimization problem, are included as Appendix 1.

Conclusion

In general, the visualization methodologies proposed here work best with a moderate number of features and a few classes due to the constraints of dimensionality and maintaining feature information. However, they have also been shown to be useful in identifying data outliers, comparing transformations, and comparing data classification complexity. With a very large number of features, the *HRC* coordinates may become more condensed due to the features being normalized. This can be mitigated in part by removing outliers, using projections, or exploring feature subsets. If performing unsupervised exploratory analysis on a dataset, a large number of classes can create a challenge unless they have break-defining feature subsets. Either way, this separation or lack of separation can still be useful information to the user.

The HRV technique itself is very simple and does not change the inherent properties of the data, thus making it very easy to interpret. Additionally, the visualization is efficient to compute. Determining optimal groupings using the objectives and formulations presented is relatively efficient, with a heuristic needed only once the number of features becomes large. In cases where the data have well-behaved class structures, the visualization provides a tool to identify this structure, and in cases where the boundaries are more complex or overlap, the visualization enables identification of such properties. If used dynamically, these visualizations can also be used for purposes of feature selection.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Todd Paciencia https://orcid.org/0000-0003-3703-1559

References

1. Leblanc J, Ward M and Wittels N. Exploring n-dimensional databases. In: *Proceedings of the 1st conference on visualization (VIS'90)*, San Francisco, CA, October 23–26, 1990, pp. 230–237. Los Alamitos: IEEE Computer Society Press.
2. Soukup T and Davidson I. *Visual data mining: techniques and tools for data visualization and mining*. Hoboken: Wiley, 2002.
3. Tufte E. *The visual display of quantitative information*. Cheshire: Graphics Press, 2001.
4. Isson JP and Harriott J. *Win with advanced business analytics*. Hoboken: John Wiley & Sons, 2013, pp.95–112.
5. Keim D. Information visualization and visual data mining. *IEEE Trans Vis Comput Graphics* 2002; 7: 100–107.
6. Grinstein G, Trutschl M and Cvek U. High-dimensional visualizations. In: *Data mining conference KDD workshop*, San Francisco, CA, January 2001, pp.7–19. New York: ACM Press.
7. Kromesch S and Juhasz S. High dimensional data visualization. In: *Proceedings of the international symposium of Hungarian researchers on computational intelligence*, Budapest, Hungary, November 2005, pp.1–12.
8. Chan W. A survey on multivariate data visualization. Technical report, Hong Kong University of Science and Technology, Department of Computer Science and Engineering, 2006.
9. Kehrer J and Hauser H. Visualization and visual analysis of multifaceted scientific data: a survey. *IEEE Trans Vis Comput Graphics* 2013; 19: 495–513.

10. Mühlbacher T, Piringer H, Gratzl S, et al. Opening the black box: strategies for increased user involvement in existing algorithm implementations. *IEEE Trans Vis Comput Graphics* 2014; 20: 1643–1652.
11. Chiu P and Bloebaum C. Hyper-radial visualization for decision-making in multi-objective optimization. In: *Proceedings of the 46th AIAA aerospace sciences meeting and exhibit*, Reno, Nevada, 7–10 January 2008, pp. 1–12. Reston: AIAA.
12. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugen* 1936; 7: 179–188.
13. LeCun Y, Cortes C and Burges C. The MNIST database, http://yann.lecun.com/exdb/mnist/ (accessed 15 January 2015).
14. Hoffman P and Grinstein G. Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. In: *NPIV workshop on new paradigms in information visualization and manipulation*, Kansas City, MO, 6 November 1999, pp.9–16. New York: ACM Press.
15. Romay D. Hyperspectral remote sensing scenes, http://www.ehu.es/ccwintco/index.php/HyperspectralRemoteSensingScenes (accessed 15 January 2015).
16. Lindsey JC, Herzberg AM and Watts DG. A method for cluster analysis based on projections and quantile-quantile plots. *Biometrics* 1987; 43: 327–341.
17. Hand DJ, Daly F, McConway K, et al. *A handbook of small data sets*. Boca Raton: CRC Press, 1993.
18. Bache K and Lichman M. UCI machine learning repository, http://archive.ics.uci.edu/ml (accessed 27 April 2015).
19. Gu Q, Li Z and Han J. Generalized Fisher score for feature selection, arxiv.org/pdf/1202.3725 (accessed 15 January 2015).
20. Lengler R and Eppler M. Towards a periodic table of visualization methods for management. In: *IASTED proceedings of the conference on graphics and visualization in engineering*, Clearwater, FL, January 3–5, 2007, pp.83–88. New York: Grove Press.
21. Inselberg A and Dimsdale B. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: *Proceedings of the first IEEE conference on visualization*, San Francisco, CA, October 23–26, 1990, pp.361–378. New York: IEEE.
22. Dang T, Wilkinson L and Anand A. Stacking graphic elements to avoid over-plotting. *IEEE Trans Vis Comput Graphics* 2010; 16: 1044–1052.
23. Xu Y, Hong W, Li X, et al. Parallel dual visualization of multidimensional multivariate data. In: *Proceedings of the IEEE international conference on integration technology*, Shenzhen, China, March 20–24, 2007, pp.263–268. New York: IEEE.
24. Keahey TA. Visualization of high-dimensional clusters using nonlinear magnification. *Electronic Imaging'99. International Society for Optics and Photonics*. 1999; 3643: 228–235.
25. Kobayashi H, Misue K and Tanaka J. Colored mosaic matrix: visualization technique for high-dimensional data. In: *IEEE international conference on information visualization*, London, UK, 16-18 July 2013, pp. 378–383. New York: IEEE.
26. Alpern B and Carter L. Hyperbox. In: *Proceedings of the IEEE conference on visualization (VIS'91)*, San Diego, CA, October 22–25, 1991, pp. 133–139. Los Alamitos, CA: IEEE Computer Society Press.
27. Rao R and Card S. The table lens: merging graphic and symbolic representations in an interactive focus + context visualization for tabular information. In: *Proceedings of the ACM CHI conference on human factors in computer systems: celebrating interdependence*, Boston, MA, April 1994, pp.318–322. New York: ACM Press.
28. Agrawal G, Lewis K and Bloebaum C. Intuitive visualization of hyperspace pareto frontier. In: *Proceedings of the 44th AIAA aerospace sciences meeting and exhibit*, Reno, Nevada, 9–12 January 2006, pp.1–11. Reston: AIAA.
29. Paciencia T. *Multi-objective optimization of mixed variable, stochastic systems using single-objective formulations*. Master's Thesis, Air Force Institute of Technology, Wright-Patterson AFB, 2008.
30. Dillon WR and Goldstein M. *Multivariate analysis: methods and applications*. New York: Wiley, 1984.
31. Johnson W and Lindenstrauss J. Extensions of Lipschitz maps into a Hilbert space. *Contemp Math* 1984; 26: 189–206.
32. Dasgupta S and Gupta A. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct Algor* 2003; 22: 60–65.
33. Achlioptas D. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J Comput Syst Sci* 2003; 66: 671–687.
34. Cox T and Cox M. *Multidimensional scaling*. 2nd ed. Boca Raton: CRC Press, 2000.
35. Bertini E, Tatu A and Keim D. Quality metrics in high-dimensional data visualization: an overview and systematization. *IEEE Trans Vis Comput Graphics* 2011; 17: 2203–2213.
36. Burer S and Letchford A. Non-convex mixed-integer nonlinear programming: a survey. *Surveys in Operations Research and Management Science* 2012; 17: 97–106.
37. Boros E and Hammer P. Pseudo-Boolean optimization. *Discrete Appl Math* 2002; 123: 155–225.
38. Yang M. Kernel eigenfaces vs. kernel fisherfaces: face recognition using kernel methods. In: *Proceedings of the IEEE international conference on automatic face and gesture recognition*, Washington, DC, 21–22 May 2002, pp.215–220. New York: IEEE.
39. Mika S, Ratsch G, Weston J, et al. Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces. *IEEE Trans Pattern Anal Mach Intell* 2003; 25: 623–628.
40. Lu J, Plataniotis K and Venetsanopoulos A. Kernel discriminant learning with application to face recognition. In: Wang L (ed) *Support vector machines: theory and applications*. Berlin: Springer, 2005, pp.275–296.
41. Duda R, Hart P and Stork D. *Pattern classification*. 2nd ed. New York: John Wiley & Sons, 2001.

42. Sain S. *Adaptive kernel density estimation*. PhD Thesis, Rice University, Houston, 1994.
43. Harrower M and Brewer C. Colorbrewer.org: an online tool for selecting colour schemes for maps. *Cartogr J* 2003; 40: 27–37.
44. Brewer C, Harrower M, Sheesley B, et al. Colorbrewer: color advice for maps, http://colorbrewer2.org/ (accessed 10 February 2018).
45. Horton P and Nakai K. A probabilistic classification system for predicting the cellular localization sites of proteins. *ISMB-96 Proc*. 1996; 4: 109–115.
46. van der Maaten LP and Hinton G. Visualizing high-dimensional data using t-SNE. *J Mach Learn Res* 2008; 9: 2579–2605.
47. Roweis S and Saul L. Nonlinear dimensionality reduction by locally linear embedding. *Science* 2000; 290: 2323–2326.

Appendix 1

This appendix includes additional comparisons of the developed HRV technique to dimension reduction methods, as well as a discussion on the optimization problem inherent to group selection during the forming of the axes.

The Wine data are depicted as projected to two dimensions by PCA and two other dimension reduction techniques in Figure 12. Local linear embedding is a technique that uses an eigen-decomposition derived from a local reconstruction of points based on nearest neighbors (here, 5-nearest).[47] t-SNE is not a strict projection, but iteratively tries to maintain similarity between points.[46] The Wine dataset includes 13 features, and is generally thought to

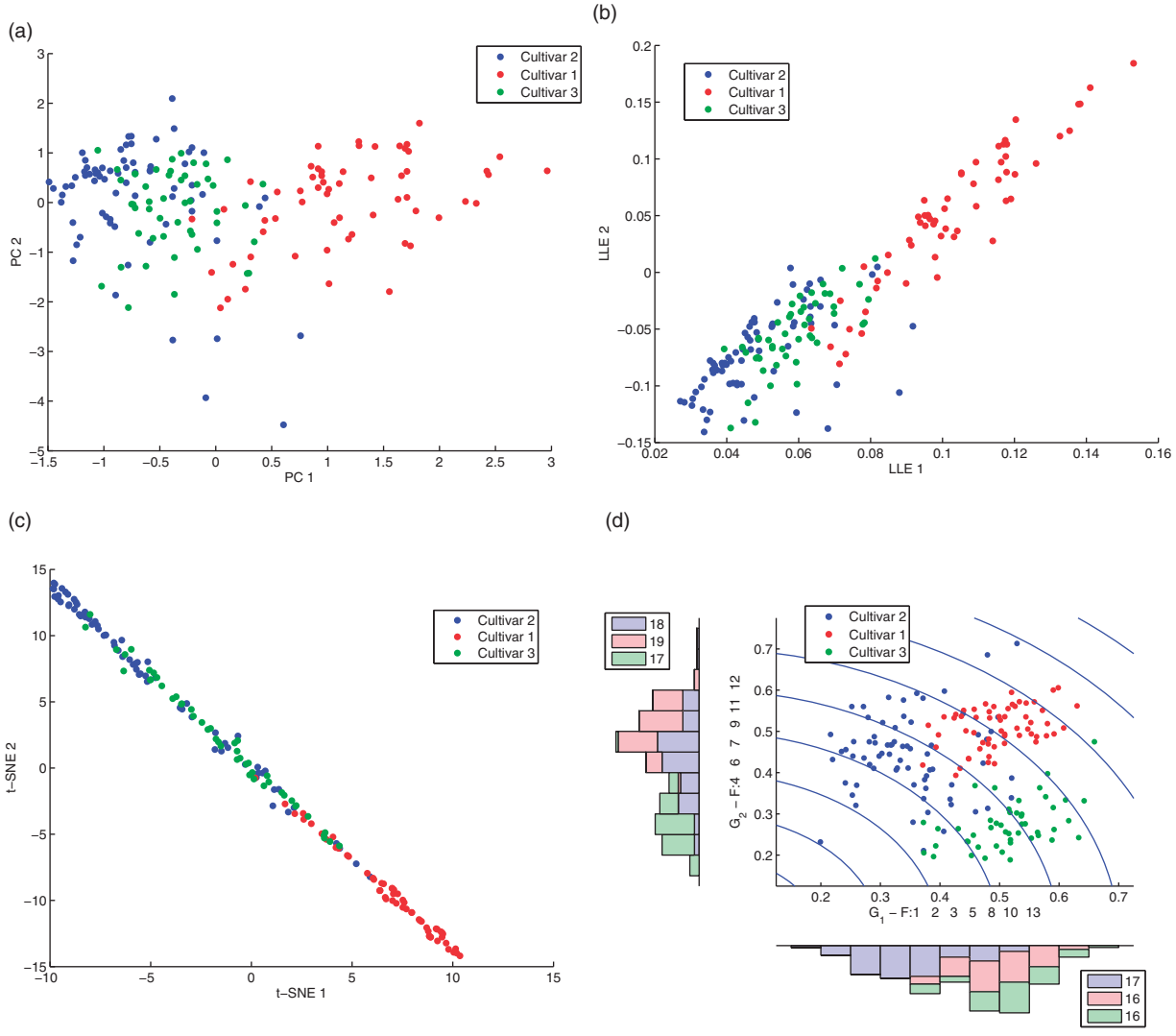

Figure 12. Wine data with two axes: (a) PCA, (b) LLE, (c) t-SNE, and (d) HRV.
HRV: hyper-radial visualization; LLE: local linear embedding; PCA: principal component analysis; t-SNE: t-Distributed Stochastic Neighborhood Embedding.

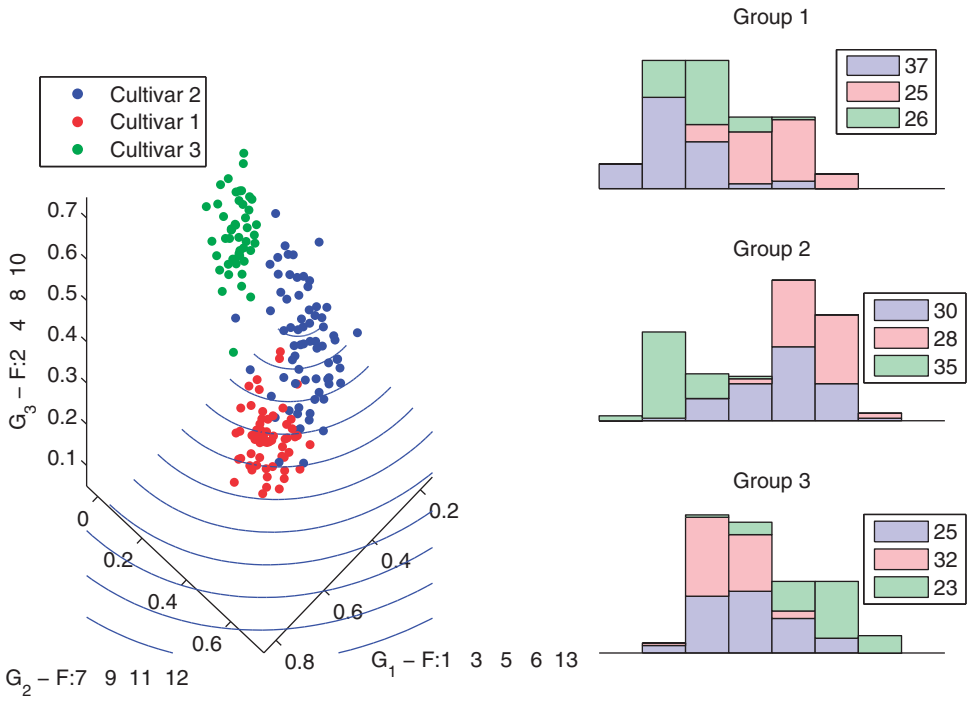

Figure 13. Wine HRV with three axes using J_1.
HRV: hyper-radial visualization.

have well-behaved class structure.[18] Figure 12(d) depicts the HRV solution, where it clearly provides the most distinction between classes, and most closely reflects differences between classes.

Figure 13 shows a three-dimensional visualization of the Wine data using J_1, representative of the solutions found using Algorithm 1 with $m = 8500$ and $q = 0.3$. The classes break out very well in this visualization. In the two-dimensional visualization, there were only $\binom{13}{7\,6} = 1716$ ways to select the groups. The three-dimensional visualization has $\binom{13}{5\,4\,4} = 90,090$ possibilities. Thus, as alluded to, a strategy other than complete enumeration is necessary to find better groupings as either the number of features or groups increases. One alternative is a very simple local search heuristic, as presented in Algorithm 1 for two groups. With three groups, the crossover adapts easily by selecting a unique feature from each group to switch.

Algorithm 1 Local search with random poll

```
Parameters: m = Max Iterations; q = Mutate Probability
i ← 1;  s ← ⌈p/2⌉;  y_1, y_2, ..., y_s ← 1;
y_{s+1}, y_{s+2}, ..., y_{s×2} ← 0;  J ← J_t(X, y)
while i < m or until convergence do
    ỹ ← y
    G_1 ← {j : y_j = 1}, G_2 ← {j : y_j = 0}
    r_1, r_2, r_3 ← random(0, 1)
    if r_3 ≥ q (Switch features between groups) then
        r_1 ← ⌈s × r_1⌉, r_2 ← ⌈s × r_2⌉
        ỹ_{G_1(r_1)} ← 0, ỹ_{G_2(r_2)} ← 1
    else (Consider random permutation)
        R_p ← Random Permutation(1 : 2s)
        ỹ_{R_p(1:s)} ← 1, ỹ_{R_p(s+1:2s)} ← 0
    end if
    J̃ ← J_t(X, ỹ)
    if J̃ > J then
        J ← J̃, y ← ỹ
    end if
    i ← i + 1
end while
```

This heuristic can be viewed as a stochastic optimization, or a very simple genetic algorithm with a single crossover. It seeks to search from the current best solution while also allowing for escape from local optima. The crossover switches one feature from each group, while for mutation, an entirely new random permutation of features in groups is used. This simple algorithm additionally serves to reduce the number of parameters and memory required for the heuristic. As with any heuristic, convergence can be dependent on the starting iterate and the number of iterations used, but this allows for some efficiency as the number of possible groupings increases dramatically with the number of features.

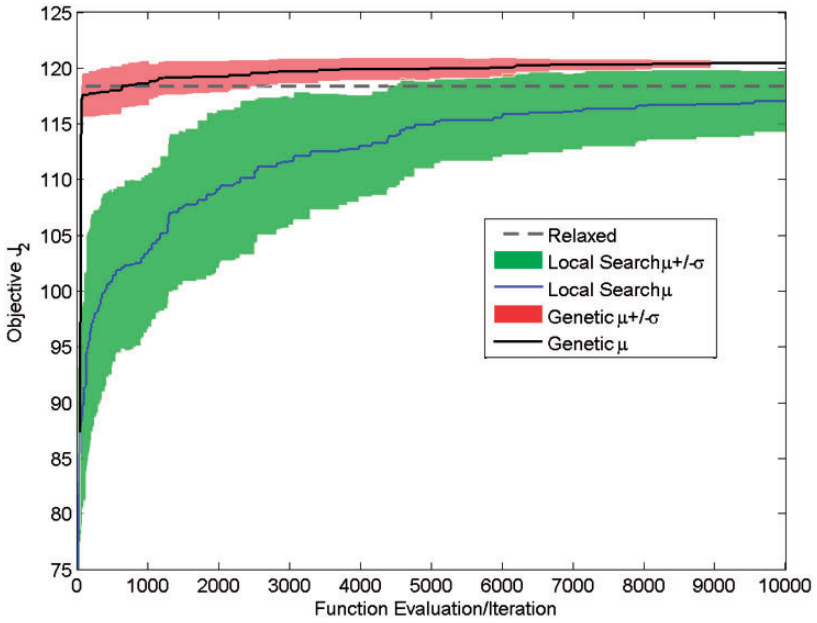

Figure 14. Wine 3-group HRV: best function value found ($\mu =$ mean, $\sigma =$ SD).
HRV: hyper-radial visualization.

As another alternative, the relaxed form of equation (6) can be solved by allowing $0 \leq y_i \leq 1$ for $i = 1, 2, \ldots, p$, rather than enforcing binary constraints during the optimization. Upon completion of an interior point method or other nonlinear programming algorithm, the variables can be set to 0 or 1 based upon their magnitude, such that the largest $\lceil p/2 \rceil$ become 1 and the remaining are set to 0.

To investigate use of Algorithm 1 against solving the relaxed problem as described, we conducted more than 30 replications per setting while varying certain parameters and solving the Wine Quality, Wine, and MNIST dataset HRV visualizations. In particular, m and q were varied for Algorithm 1. Across the heuristic and relaxed problems, we also varied the objective and number of groups. In general, Algorithm 1 showed better objective values when q was non-zero, and as m increased (for obvious reasons), was fairly efficient in converging to local maxima, and showed the ability to outperform the relaxed problem. The interior point method for the relaxed problem also proved to be highly efficient.

Comparison of methods is somewhat hard to show due to the true optima being unknown, but Figure 14 depicts the objective function value convergence of Algorithm 1 in solving the Wine dataset with J_2. Also shown are the solution found using an interior point method on the relaxed problem, and a genetic algorithm. For the genetic algorithm, mutation was as in Algorithm 1 with probability 0.2, and a single-feature switch was done between randomly chosen parents in a population of 50 solutions. Over 30 replications, both the genetic algorithm and the local search converged to an equal or better solution than the relaxed problem, on average. The diversity within the genetic algorithm allowed it to converge faster, but we found that with larger-dimension datasets this benefit was somewhat negated by the much larger memory required to store the population. On the MNIST dataset using three groups, the relaxed problem often got stuck at the initial solution, while Algorithm 1 improved from the initial solution in a similar fashion to the curve shown in Figure 14.