Structural Manifold Analysis: A New Method for Dimensionality Reduction Based on a Computational Theory of Human Concept Learning

Ronaldo Vigo
Center for the Advancement of Cognitive Science
Psychology Department
Ohio University
Athens, OH. USA
vigo@ohio.edu

Basawaraj
Center for the Advancement of Cognitive Science
School of Electrical Engineering and Computer Science
Ohio University
Athens, OH. USA
basawaraj.basawaraj.1@ohio.edu

A new non-parametric method for reducing the number of dimensions in binary and continuous data, and for measuring the complexity of binary and continuous datasets, is introduced. The method, named Structural Manifold Analysis (SMA), is based on "Generalized Invariance Structure Theory" [5-6], a theory that has been successful in characterizing and accurately predicting human concept learning and categorization performance. SMA is unique among data reduction and classification methods because it determines the degree of "diagnosticity" or classification potential of each of the dimensions in a multidimensional dataset up to a pre-specified discrimination resolution threshold. We compared SMA to five classical and recent dimensionality reductions methods using supervised and unsupervised classification techniques on 304 actual and simulated datasets. Overall, SMA performed as well as or better than any of the five methods tested while providing significant advantages over each.

Keywords— dimensionality reduction; human classification; machine classification; machine learning

I. INTRODUCTION

Data sets in domains such as machine learning, data mining and numerical analysis have high dimensional spaces with tens, hundreds or thousands of dimensions. When analyzing high dimensional data various problems, issues that do not arise in the three-dimensional physical space, occur. Many applications, such as classification, can require high dimensional data for reliable results. But as the dimensionality increases the volume of the data space increases but the data itself becomes sparse. In addition, many applications, such as classification, depend on detecting areas where objects with similar properties group together; but with high dimensional space general data organization strategies are not always efficient. The problems occur because the data becomes sparse as the dimensionality increases and efficiency of common data organization techniques decreases.

The problems with analysis of high dimensional data can occur because the algorithms do not scale well to high dimensional data. Solutions include either changing the algorithm or preprocessing data into a lower dimensional space. Various dimensionality reduction techniques have been proposed in literature but these techniques can generally be classified as either feature extraction or feature selection techniques. Note that in this work the terms features and dimensions are used interchangeably and mean the same unless specified otherwise. Feature extraction techniques assume that the data of interest lie on an embedded manifold within the higher dimensional space and transform the original set of features into a reduced set of features. The transformation can either be linear or non-linear and involves combining existing features to create new features. Principal component analysis (PCA) [7], nonlinear PCA (NLPCA) [8], kernel PCA (KPCA) [9], and singular value decomposition (SVD) [10] are some of the approaches to feature extraction. Unlike feature extraction techniques the feature selection techniques try to select a subset of the original features and are widely used in pattern recognition tasks to identify the characteristic features of a given dataset. Minimum-redundancy-maximum-relevance (mRMR) [11] is an example of feature selection technique. The best subset of features for optimal characterization, minimum classification error, is always obtained by using an exhaustive search, but it is computationally intensive and is not always feasible.

II. A THEORY OF HUMAN CLASSIFICATION PERFORMANCE

Generalized Invariance Structure Theory (GIST) is a mathematical theory of human concept learning that has been successful in characterizing and in making accurate predictions with respect to human categorization/classification performance [5-6]. The concept learning law at the core of GIST, with either a single scaling parameter or without parameters, accounts for nearly all of the variance in the data from several large scale studies on human classification performance [2, 5, 6]. The core idea underlying GIST is that humans detect atomic invariance patterns (named "categorical invariants") in sets of objects (i.e., categorical stimuli). The human conceptual system then computes the proportions of detected categorical invariants (with respect to each dimension) to the number of objects in the categorical stimulus. This structural information is then used to determine

the degree of diagnosticity (on an inverse scale) of each relevant dimension of the categorical stimulus – or, in other words, the extent to which each of the dimensions comprising any categorical stimulus is able to predict object membership in the categorical stimulus. This information is then used by observers to form classification rules by discarding redundant dimensions and keeping as many as possible (as determined by processing limits) of the diagnostic ones. The degree of perceived difficulty or subjective complexity of a categorical stimulus is then characterized by the law of invariance (also known as the Generalized Invariance Structure Theory Model or GISTM) which states that the degree of learning difficulty of a set of objects defined dimensionally is directly proportional to its cardinality and inversely proportional to the exponent of its overall degree of invariance [2, 5, 6]. We use these ideas from GIST to develop a new and effective method for multivariate data analysis that conforms to human intuitions. We also aim to determine how effective a theory that captures accurately human classification performance can capture the structure of multidimensional datasets in terms of a reduced set of dimensions.

III. CATEGORICAL INVARIANCE

The formal description of the process specified in GIST is based on a notion referred to as *categorical invariance*, and on several other related measures, first introduced in [1-6]. Here we mention only three. The first measure, referred to as "*partial categorical invariance*", is a measure of the degree of local or partial relational homogeneity of a dimensionally-defined set of objects with respect to a particular dimension. The second measure, referred to as "degree of categorical invariance", is a measure of the overall or global degree of relational homogeneity of a dimensionally-defined set of objects with respect to all of its relevant dimensions. The third measure is a measure of the perceived structural complexity of a dimensionally-defined set of objects based on its cardinality (i.e., its size) and its degree of categorical invariance. The strength of these measures stems from their ability to apply to both binary and continuous multidimensional data.

Structural Manifold Analysis (SMA) consists of the application of these measures from GIST (and a dimension selection heuristic introduced in section V) on multivariate data sets and not on the categorical stimuli they were intended for. Indeed, in principle, the representation of a data sets in terms of a matrix of variables is no different to the representation of a categorical stimulus in GIST in terms of a matrix of dimensional values; thus, GIST generalizes seamlessly to both cases. In section VI, we compare SMA to five classic and recent dimensionality reduction methods and test these using five prominent classification techniques.

To understand how categorical invariance measures the homogeneity of datasets, we first perturbed the dataset I in Table I with respect to the binary dimension of protein-content (the third dimension). We do this by assigning the opposite protein-content value to each of the four food brands in the set. This yields the perturbed dataset $\{(1,1,0,0), (1,1,1,1), (1,1,1,0), (1,1,0,1)\}$ which indicates a transformation of the dataset {(1,1,1,0), (1,1,0,1), (1,1,0,0), (1,1,1,1)} along its third dimension as shown in Table III. More specifically, all ones become zeros and all zeros become ones for the third value of each vector representing a data point in the set.

In general, let M be a multidimensional data set and let $T_i(\mathrm{M})$ stand for the result of applying such a transformation T_i on the multidimensional dataset M along its *i*-th dimension. After the transformation has been applied, we compare the original data to the perturbed data and recognize they have four points in common with respect to the protein-content dimension. Thus, four out of the four data points remain the same. This ratio is a measure of the partial homogeneity of the data with respect to the dimension of protein-content in our example above. The ratio can be written in general terms as a lambda operator on M:

$$\Lambda_i(\mathrm{M}) = \frac{\left|\mathrm{M} \cap T_i(\mathrm{M})\right|}{\left|\mathrm{M}\right|} \tag{1}$$

In (1), if M is a binary dataset of D dimensions $(D \geq 1)$, then, for any dimension i $(1 \leq i \leq D)$, the transformation T_i on M is defined as follows: $T_i(\mathrm{M}) = \{(x_1,\ldots,x_i',\ldots,x_D) \,|\, (x_1,\ldots,x_i,\ldots,x_D) \in \mathrm{M}\}$ where $x_i' = 1$ if $x_i = 0$ and $x_i' = 0$ if $x_i = 1$. Furthermore, $\left|\mathrm{M} \cap T_i(\mathrm{M})\right|$ stands for the number of points that the transformed dataset M has in common with M with respect to the transformation T_i. Doing this for each of the dimensions, all of the local (partial) homogeneities of the dataset are generated. Table III illustrates this transformative process for each dimension D1 through D4 on the first and second datasets of four food brands (i.e., Data I and Data II). It also shows the final set of points that remain the same in the data after their transformation. Note that compound transformations (i.e., involving 2 or more dimensions) are possible and were considered by the author of GIST when it was first developed. The result was that, from a cognitive standpoint, one gained little to nothing in terms of accounting for human classification data. In addition, the theory and its underlying models are not nearly as parsimonious, becoming instead post-hoc and arbitrary in character. Nevertheless, strictly for the purpose of data analysis, we acknowledge that this type of expanded analysis may merit further investigation in future work.

Now that we have explained the nature of categorical invariance, we are prepared to define the logical manifold operator Λ on a multivariate (i.e., multidimensional) dataset M (where $D \geq 1$ is the number of dimensions in M) as follows:

$$\Lambda(\mathrm{M}) = \left(\frac{\left|\mathrm{M} \cap T_1(\mathrm{M})\right|}{\left|\mathrm{M}\right|}, \frac{\left|\mathrm{M} \cap T_2(\mathrm{M})\right|}{\left|\mathrm{M}\right|}, \ldots, \frac{\left|\mathrm{M} \cap T_D(\mathrm{M})\right|}{\left|\mathrm{M}\right|} \right) \tag{2}$$

TABLE I. DATA SET I WITH FOUR DIMENSIONS

Brand	Fat	Sugar	Protein	Fiber
A	1	1	1	0
B	1	1	0	1
C	1	1	0	0
D	1	1	1	1

TABLE II. DATA SET II WITH FOUR DIMENSIONS

Brand	Fat	Sugar	Protein	Fiber
E	0	1	1	1
F	0	0	1	1
G	0	0	0	0
H	1	0	1	1

TABLE III. INVARIANCE OF MULTIDIMENSIONAL DATA OF SETS I AND II IN TABLES I AND II RESPECTIVELY IN RESPECT TO 4 DIMENSIONS

	Data I	**Transformed Data I**	**Set Intersection**
D1	{1110, 1101, 1100, 1111}	{0110, 0101, 0100, 0111}	∅
D2	{1110, 1101, 1100, 1111}	{1010, 1001, 1000, 1011}	∅
D3	{1110, 1101, 1100, 1111}	{1100, 1111, 1110, 1101}	{1100, 1111, 1110, 1101}
D4	{1110, 1101, 1100, 1111}	{1111, 1100, 1101, 1110}	{1111, 1100, 1101, 1110}

	Data II	**Transformed Data II**	**Set Intersection**
D1	{0111, 0011, 0000, 1011}	{1111, 1011, 1000, 0011}	{1011, 0011}
D2	{0111, 0011, 0000, 1011}	{0011, 0111, 0100, 1111}	{0011,0111}
D3	{0111, 0011, 0000, 1011}	{0101, 0001, 0010, 1001}	∅
D4	{0111, 0011, 0000, 1011}	{0110, 0010, 0001, 1010}	∅

The vector of the local homogeneities generated by the Λ operator is referred to in GIST as the logical manifold of the set of points M. Accordingly, the degrees of global homogeneity of any multidimensional dataset may then be measured by computing the Euclidean distance between its logical manifold and the **0** = (0,…,0) logical manifold of the same dimensionality. The **0** manifold represents total absence of invariance with respect to each dimension of a multivariate dataset. In other words, all the local homogeneities of a multidimensional dataset with a **0** logical manifold are equal to 0 (a situation where all the dimensions are fully diagnostic or non-reducible or for which there are no dimensional redundancies). Thus, the **0** logical manifold is a baseline or lower bound for the degree of global invariance of the set of data points. The distance from this zero point in "dataset space" defines the relative degree of overall invariance or homogeneity Φ of any multivariate dataset M as shown in (3). Simply stated, Φ is a measure of the compressibility of M. Note that this equation is based on the Euclidean distance measure but it can be generalized to the Minkowski distance measure as done in [2, 5, 6] and in the next section.

$$\Phi(\mathrm{M}) = \left[\sum_{i=1}^{D} \left[\frac{\left| \mathrm{M} \cap T_i(\mathrm{M}) \right|}{\left| \mathrm{M} \right|} \right]^2 \right]^{1/2} \tag{3}$$

IV. STRUCTURAL MANIFOLDS AND DATA COMPLEXITY

One of the main accomplishments of GIST is its generalization from the notion of a logical manifold to a structural manifold. The latter can handle continuous and binary dimensions. To recognize the role that similarity plays in the determination of the degree of local relative homogeneity of a multidimensional dataset we turn to the concept of symmetry. For example, consider the partial symmetry shown in Fig. 1 between two 3-dimensional data points from a 3-dimensional dataset. With respect to the first dimension (i.e., when the first dimension is disregarded) both data points are identical. This idea is referred to in GIST as the *invariance-similarity principle* and is used to develop a general theory of conceptual behavior. Here, we shall use it to generalize the logical manifold operator given in (1) so that it applies to both binary and continuous domains. Under such generalization, the *logical manifold operator* will be called the *structural manifold operator*; accordingly, the manifolds generated by the structural manifold operator shall be referred to as *structural manifolds*. The core idea is that the degree of redundancy and, hence, (on an inverse scale) diagnosticity of a dimension is revealed by how much its temporary removal or suppression has on the homogeneity of the dataset as a whole. In GIST, this suppression of a dimension is referred to as *binding* the dimension. Next, we shall introduce a similarity measure that formalizes this basic idea.

In so doing, we shall employee the following additional notation: let X be a multidimensional dataset and $|\mathrm{X}|$ stand for the cardinality (i.e., the number of elements) of X. Let the points in X be represented by the vectors $\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2, \ldots, \vec{\mathbf{x}}_n$ (where $n = |\mathrm{X}|$) and let the vector $\vec{\mathbf{x}}_j = (x_1, \ldots, x_D)$ be the *j*-th *D*-dimensional data point in X (where *D* is the number of dimensions or variables of the dataset). Furthermore, let $\vec{\mathbf{x}}_{ji}$ be the value of the *i*-th dimension of the *j*-th point in X. We shall assume throughout our discussion that all dimensional values are real numbers greater than or equal to zero. Finally, let $S(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k)$ stand for the similarity of data point $\vec{\mathbf{x}}_j \in \mathrm{X}$ to data point $\vec{\mathbf{x}}_k \in \mathrm{X}$.

Our aim in this section is to introduce a generalized version of the structural manifold operator of (1) using the invariance-similarity principle. To do this, we use the generalized Euclidean distance operator Δ^r (a.k.a. *Minkowski*

distance) between two data points $\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k \in \mathrm{X}$ defined as follows:

101	→	001		
			←→	101=001
001	→	101		**(***identical with respect to* **D1)**

Fig. 1. Equivalence of Categorical Invariance (with respect to D1) to the Partial Similarity between Two Points

$$\Delta^r(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k) = \left[\sum_{i=1}^{D} \left| \vec{\mathbf{x}}_{ji} - \vec{\mathbf{x}}_{ki} \right|^r \right]^{1/r} \tag{4}$$

Furthermore, we introduce a new kind of distance operator termed the *partial distance operator* $\Delta^r_{[d]}$:

$$\begin{aligned} \Delta^r_{[d]}(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k) &= \left[\sum_{i \neq d} \left| \vec{\mathbf{x}}_{ji} - \vec{\mathbf{x}}_{ki} \right|^r \right]^{1/r} \\ &= \sqrt[r]{\left[\sum_{i=1}^{D} \left| \vec{\mathbf{x}}_{ji} - \vec{\mathbf{x}}_{ki} \right|^r \right] - \left[\left| \vec{\mathbf{x}}_{jd} - \vec{\mathbf{x}}_{kd} \right|^r \right]} \end{aligned} \tag{5}$$

Equation (5) simply takes the sum of the differences between two data points in X across each of their dimensional values except the value corresponding to the "bound" *d-th* dimension ($1 \leq d \leq D$). In other words, it computes the partial distance between any two data points $\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k \in \mathrm{X}$, by excluding dimension *d* in the computation of the Minkowski generalized metric. For example, for a dataset consisting of four points, we can conveniently represent these partial pairwise distances with respect to a dimension *d* with the following partial distances matrix:

$$\mathbf{D}^r_{[d]}(\mathrm{X}) = \begin{bmatrix} \Delta^r_{[d]}(\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_1) & \Delta^r_{[d]}(\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2) & \Delta^r_{[d]}(\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_3) & \Delta^r_{[d]}(\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_4) \\ \Delta^r_{[d]}(\vec{\mathbf{x}}_2, \vec{\mathbf{x}}_1) & \Delta^r_{[d]}(\vec{\mathbf{x}}_2, \vec{\mathbf{x}}_2) & \Delta^r_{[d]}(\vec{\mathbf{x}}_2, \vec{\mathbf{x}}_3) & \Delta^r_{[d]}(\vec{\mathbf{x}}_2, \vec{\mathbf{x}}_4) \\ \Delta^r_{[d]}(\vec{\mathbf{x}}_3, \vec{\mathbf{x}}_1) & \Delta^r_{[d]}(\vec{\mathbf{x}}_3, \vec{\mathbf{x}}_2) & \Delta^r_{[d]}(\vec{\mathbf{x}}_3, \vec{\mathbf{x}}_3) & \Delta^r_{[d]}(\vec{\mathbf{x}}_3, \vec{\mathbf{x}}_4) \\ \Delta^r_{[d]}(\vec{\mathbf{x}}_4, \vec{\mathbf{x}}_1) & \Delta^r_{[d]}(\vec{\mathbf{x}}_4, \vec{\mathbf{x}}_2) & \Delta^r_{[d]}(\vec{\mathbf{x}}_4, \vec{\mathbf{x}}_3) & \Delta^r_{[d]}(\vec{\mathbf{x}}_4, \vec{\mathbf{x}}_4) \end{bmatrix} \tag{6}$$

And more generally as:

$$\mathbf{D}^r_{[d]}(\mathrm{X}) = \begin{bmatrix} \Delta^r_{[d]}(\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_1) & \cdots & \Delta^r_{[d]}(\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_P) \\ \vdots & \ddots & \vdots \\ \Delta^r_{[d]}(\vec{\mathbf{x}}_P, \vec{\mathbf{x}}_1) & \cdots & \Delta^r_{[d]}(\vec{\mathbf{x}}_P, \vec{\mathbf{x}}_P) \end{bmatrix} \tag{7}$$

Similarly, we can define the partial similarity between two data points as is done in multidimensional scaling theory [12] as a monotonically decreasing function *F* of the partial distance between the two data points.

$$S_{[d]}(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k) = F(\Delta^r_{[d]}(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k)) \tag{8}$$

The simplest non-trivial such function is the additive inverse of the standardized partial distance measure $\Delta^r_{[d]}(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k)$ using the city block metric (i.e., when *r*=1) as shown in (9). The standardization is achieved by a linear transformation into the interval [0, 1] as seen in (10) where the *max* and *min* of a matrix are respectively its largest and smallest element and the $\mathtt{max}(\mathbf{D}^r_{[d]}(\mathrm{X}) \neq \mathtt{min}(\mathbf{D}^r_{[d]}(\mathrm{X}))$ for any *d* and *r*.

$$S_{[d]}(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k) = 1 - \omega(\Delta^1_{[d]}(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k)) \tag{9}$$

$$\omega(\Delta^1_{[d]}(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k)) = \frac{\Delta^1_{[d]}(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k) - \mathtt{min}(\mathbf{D}^1_{[d]}(\mathrm{X}))}{\mathtt{max}(\mathbf{D}^1_{[d]}(\mathrm{X})) - \mathtt{min}(\mathbf{D}^1_{[d]}(\mathrm{X}))} \tag{10}$$

Although other proven measures of similarity as functions of distance may be used, such as the exponential function $e^{-\Delta^r_{[d]}(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k)}$, to simplify our explanation and analysis, the similarity measure in (9) will do for the remainder of this article. Either measure may be used depending on whether the data is human subject data (exponential) or objective data (simple inverse). Now we can construct the matrix of the pairwise partial similarities in a dimensional dataset. Note that the example given in (11) below consists of a 4 dimensional dataset with four data points as those shown in Table III. By convention, we have excluded reflexive or self-similarities in the diagonal of the matrix. However, we include symmetric comparisons since they are at the heart of the operator introduced in (5) (the importance of including symmetric comparisons can be seen Fig. 1), and our goal is to be faithful to this invariance measure.

$$\mathbf{S}_{[d]}(\mathrm{X}) = \begin{bmatrix} - & S_{[d]}(\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2) & S_{[d]}(\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_3) & S_{[d]}(\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_4) \\ S_{[d]}(\vec{\mathbf{x}}_2, \vec{\mathbf{x}}_1) & - & S_{[d]}(\vec{\mathbf{x}}_2, \vec{\mathbf{x}}_3) & S_{[d]}(\vec{\mathbf{x}}_2, \vec{\mathbf{x}}_4) \\ S_{[d]}(\vec{\mathbf{x}}_3, \vec{\mathbf{x}}_1) & S_{[d]}(\vec{\mathbf{x}}_3, \vec{\mathbf{x}}_2) & - & S_{[d]}(\vec{\mathbf{x}}_3, \vec{\mathbf{x}}_4) \\ S_{[d]}(\vec{\mathbf{x}}_4, \vec{\mathbf{x}}_1) & S_{[d]}(\vec{\mathbf{x}}_4, \vec{\mathbf{x}}_2) & S_{[d]}(\vec{\mathbf{x}}_4, \vec{\mathbf{x}}_3) & - \end{bmatrix} \tag{11}$$

The values on the diagonal of the above matrix are equal to one (since the partial distance of any stimulus to itself is zero) but these do not play a role in estimating the overall local homogeneity of the dataset X. Adding the values of the similarity matrix that correspond to differences within a chosen *distance resolution threshold* for each dimension *d* we can get the following expression which is functionally analogous to the local homogeneity operator given in (5) (for any pair of objects $(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k)$ where $\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k \in \mathrm{X}$, $j \neq k$, and $j, k \in \{1, 2, ..., |\mathrm{X}|\}$):

$$H_{[d]}(\mathrm{X}) = \frac{\sum_{0 \leq \Delta^r_{[d]}(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k) \leq \tau_d} S_{[d]}(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k)}{|\mathrm{X}|} \tag{12}$$

The equation above defines the local homogeneity $H_{[d]}$ of a *D*-dimensional dataset X with respect to dimension *d*. $H_{[d]}(\mathrm{X})$ is the ratio between: 1) the sum of the similarities in the matrix $\mathbf{S}^{\mathrm{X}}_{[d]}$ (for a particular bound dimension *d*) that correspond to distances in the $[0, \tau_d]$ distance resolution interval, and 2) the number of items in the dataset X. When the partial distances are close to zero, the points are, for all intent and purpose, treated as identical. Equation 13 below shows the matrix used to calculate the local homogeneity (with respect to dimension 3) of the dataset $\mathrm{A} = \{1110,\ 1101,\ 1100,\ 1111\}$ (this is the first dataset depicted in Table I) when we let $\tau_1 = 0$ and *r*=1.

$$\mathbf{S}_{[3]}(\mathrm{A}) = \begin{bmatrix} - & S_{[3]}(\vec{\mathbf{x}}_1,\vec{\mathbf{x}}_2) & S_{[3]}(\vec{\mathbf{x}}_1,\vec{\mathbf{x}}_3) & S_{[3]}(\vec{\mathbf{x}}_1,\vec{\mathbf{x}}_4) \\ S_{[3]}(\vec{\mathbf{x}}_2,\vec{\mathbf{x}}_1) & - & S_{[3]}(\vec{\mathbf{x}}_2,\vec{\mathbf{x}}_3) & S_{[3]}(\vec{\mathbf{x}}_2,\vec{\mathbf{x}}_4) \\ S_{[3]}(\vec{\mathbf{x}}_3,\vec{\mathbf{x}}_1) & S_{[3]}(\vec{\mathbf{x}}_3,\vec{\mathbf{x}}_2) & - & S_{[3]}(\vec{\mathbf{x}}_3,\vec{\mathbf{x}}_4) \\ S_{[3]}(\vec{\mathbf{x}}_4,\vec{\mathbf{x}}_1) & S_{[3]}(\vec{\mathbf{x}}_4,\vec{\mathbf{x}}_2) & S_{[3]}(\vec{\mathbf{x}}_4,\vec{\mathbf{x}}_3) & - \end{bmatrix} \quad (13)$$
$$= \begin{bmatrix} - & 0 & 1 & 0 \\ 0 & - & 0 & 1 \\ 1 & 0 & - & 0 \\ 0 & 1 & 0 & - \end{bmatrix}$$

Note that the computed matrix in (13) contains 4 ones that represent four identical pairs of data points. Applying (12) we get

$$H_{[1]}(\mathrm{X}) = \frac{\sum_{0 \le \Delta^1_1(\vec{\mathbf{x}}_j,\vec{\mathbf{x}}_k) \le 0} S_{[1]}(\vec{\mathbf{x}}_j,\vec{\mathbf{x}}_k)}{|\mathrm{X}|} = \frac{1+1+1+1}{4} = 1 \quad (14)$$

Lastly, we define the generalized structural manifold with (15). This construct is analogous to the one defined in (2), except that it applies to both binary and continuous dimensions and is equipped with a distance resolution threshold.

$$\Lambda(\mathrm{X}) = \left(H_{[d=1]}(\mathrm{X}), H_{[d=2]}(\mathrm{X}), ..., H_{[d=D]}(\mathrm{X})\right) \quad (15)$$

We can also specify the particular partial or local homogeneity of X (comprising the structural manifold) as seen in the equation below.

$$\Lambda^{\tau_d}_{d}(\mathrm{X}) = \frac{\sum_{0 \le \Delta^r_{[d]}(\vec{\mathbf{x}}_j,\vec{\mathbf{x}}_k) \le \tau_d} S_{[d]}(\vec{\mathbf{x}}_j,\vec{\mathbf{x}}_k)}{|\mathrm{X}|} \quad (16)$$

With (16) (when r=1 and $\tau_d = 0$ for all d) we can compute the partial homogeneities of the multidimensional datasets from Table III and get results consistent with those shown in the third column of the table. Furthermore, using our new formulation of the generalized structural manifold operator, we can compute the structural complexity for any multivariate dataset X defined over $D \ge 1$ dimensions and for any pair of objects $(\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k)$ (such that $\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_k \in \mathrm{X}$, $j \ne k$, $j,k \in \{1,2,...,|\mathrm{X}|\}$, using Vigo's law of invariance [5, 6] as in (17). Although this exponential function of invariance is more accurate from a human performance perspective, the identity +1 function (18) achieves good approximations and is computationally more parsimonious.

$$\psi(\mathrm{X}) = |\mathrm{X}| \cdot e^{-k\left[\left[\sum_{d=1}^{D}\left[H_{[d]}(\mathrm{X})\right]^2\right]^{\frac{1}{2}}\right]^2} = |\mathrm{X}| \cdot e^{-k\hat{\Phi}^2(\mathrm{X})} \quad (17)$$

$$\psi(\mathrm{X}) = \frac{|\mathrm{X}|}{k\left[\sum_{d=1}^{D}\left[H_{[d]}(\mathrm{X})\right]^2\right]+1} = \frac{|\mathrm{X}|}{k\hat{\Phi}^2(\mathrm{X})+1} \quad (18)$$

A Matlab program (ver. 7) to compute the structural manifold and structural complexity of multidimensional dataset is available at http://www.scopelab.net/programs.htm.

V. SELECTING DIMENSIONS FROM STRUCTURAL MANIFOLDS

We extracted the two structural manifolds along the lines of the dependent variable values (e.g., whether a tumor is cancerous or not). However, in general this is not necessary and one may extract the structural manifold of an unpartitioned dataset. We apply a heuristic procedure on the components of each structural manifold in order to eliminate the most redundant dimensions.

The elements of each of the two structural manifolds (with their dimensional labels as shown in TABLE III) are ordered from smallest to largest and, whenever there are identical values in the set, these are ordered in ascending order of their dimensional labels: for example, suppose that the same degree of local homogeneity .5 is associated with dimensions D3, D1, and D4, then these three labels are arranged in ascending dimensional order within the two structural manifolds: D1(.5), D3(.5), D4(.5). The relatively less diagnostic dimensions from the two structural manifolds are removed to produce two sparse sets of dimensions. To do this, we delete dimensions with local homogeneity greater than the median[1] of the values in each of the two structural manifolds. One of these sparse sets will be the "base" set and the other set is referred to as the "target" set. The base set is the set containing the smallest value (i.e., the most diagnostic dimension) among all the values in the two sets. If the smallest value is shared by both sets, then the set with the next smallest value is the base set, and so on. On the other hand, if the sets are identical, then either can serve as a base set.

We decide on the number of desired diagnostic dimensions. Next we determine whether the first dimensional label of the base set is in the target set. If so, then it is the first candidate diagnostic dimension. In either case, we proceed to

[1]Although we employ the median in our heuristic, one can use a more aggressive rule of thumb based on the first quartile – or even smaller cutoffs – depending on the desired level of reduction by the user and on whether the number of dimensions or characteristics involved are large.

the next dimensional label in the base set and, once again, determine membership in the target set. This search procedure is repeated until the desired number of diagnostic dimensions are obtained. If this result is an empty set, no diagnostic reduction is possible.

If the resultant dimensional reduction is not as low as desired, we repeat the process using a higher discrimination level when computing the structural manifolds. Note that we begin by setting it to 0, this provides the most stringent test for the presence of invariance structure in the data set but assumes identity of points which is not often a realistic psychological assumption. If this does not work, increase the value to .05 and .1. The use of .05 value increment is more realistic and is based on the psychologically meaningfully idea that humans can discriminate values in a 1-20 subjective judgment scale. This value has worked well for us in terms of relaxing the similarity criterion and henceforth finding hidden invariance structure in the data without relaxing it too much. Non-psychological arguments for setting tau to different values will be explored in future research. In general, the discrimination level should not be higher than necessary in order to determine the desired number of diagnostic dimensions. Here is the pseudo-code for this heuristic procedure:

Let SM_1 and SM_2 be the structural manifolds and TS the target set.
1. Sort SM_1 and SM_2
SM_1 = sort(SM_1, 'ascend');
SM_2 = sort(SM_2, 'ascend');
2. If SM_1(i) > median(SM_1) then SM_1(i) = [];
Similarly if SM_2(i) > median(SM_2) then SM_2(i) = [];
3. Structural manifold with minimum value not shared with other is the Base Set (BS)
4. Set number of diagnostic dimensions = N;
5. Diagnostic Dimensions, DD = unsorted_intersect(BS , TS);
6. IF DD = {}, no reduction possible. Go to (8).
7. IF length(DD) < N, repeat (1) to (6) using lower discrimination level
ELSE DD(N+1:end) = [];
8. STOP

VI. EXPERIMENTS

We tested our approach on three datasets and a variety of simulated datasets based on these three datasets. For these datasets we compare performance of SMA with three feature reduction (PCA, NLPCA and KPCA) and two feature selection (mRMR and exhaustive) techniques. The performance of dimensionality reduction techniques was tested using three classification and two clustering techniques. Linear discriminant analysis (LDA) [13], k-nearest neighbor (KNN) [14] and support vector machine (SVM) [9] are the classification techniques used while k-means [13] and hierarchical [13] are the clustering techniques used. LDA was performed assuming prior probabilities were uniformly distributed and a pooled covariance matrix was estimated from the training data. KNN classifier used Euclidean distance and samples were assigned to the class of the majority of the k nearest neighbors, where k=1. SVM was implemented using sequential minimum optimization (SMO) algorithm with linear kernel, soft margin C=1 and 5% of variables allowed to violate Karush-Kuhn-Tucker (KKT) conditions. K-means and hierarchical clustering used squared Euclidean and Euclidean distance measures respectively with number of clusters set equal to two, the number of classes in test data. K-means clustering was repeated five times using new initial cluster centroid positions and the solution with the lowest within-cluster sums of point-to-centroid distances was used. Hierarchical clustering used the complete-linkage agglomerative approach.

In sub-section A we provide a brief introduction of the datasets and in sub-section B we compare SMA against the five dimensionality reduction techniques mentioned above. The results show the advantages of our SMA approach over the other approaches.

A. Datasets

Wisconsin Diagnostic Breast Cancer (WDBC) [15], 1984 United States Congressional Voting Records (CVRD) [15] and Alzheimer's [16] were the three data sets used to compare performance of SMA against the five other dimensionality reduction techniques. WDBC dataset consists of 569 patient samples, with 357 patients diagnosed benign and 212 diagnosed malignant. The actual WDBC dataset consists of 10 real-valued features with the mean, standard error, and "worst" or largest (mean of the three largest values) of these features computed for each image, resulting in 30 features. In this work only the mean values were used, hence the dataset used only had 10 dimensions.

The original CVRD dataset has 435 samples, classified as either democrat (267 samples) or republican (168 samples), with 16 dimensions representing 16 key votes identified by Congressional Quarterly Almanac (QOA). Since the original dataset had missing values we discarded instances that had any missing attribute (dimension) and ended with 232 instances divided as follows – 108 Republicans and 124 Democrats. The third dataset is the memory test database on Alzheimer's and normal patients with 6 real-valued dimensions and available from the R project and is part of the Independent Factor Analysis (IFA) package. We use a subset of the data set that has 31 instances divided into two classes: Alzheimer's (15) and Normal (16).

From the CVRD dataset we generated 30 simulated datasets with the same number of instances using the same range of values found in their actual counterparts. The binary values for each dimensional variable were sampled at random using the random number generator in Excel 2010. For the WDBC and Alzheimer's datasets we simulated 120 datasets by sampling each dimensional variable from a Gaussian (30 datasets), exponential (30 datasets), and uniform distribution (30 datasets). A fourth type of dataset was simulated by sampling from all three distributions at random per dimensional variable (30 datasets). In short, we tested a total of 273 datasets (31 with dichotomous dimensions and 242 with continuous dimensions) using SMA and the following four aforementioned classification methods. Table IV illustrates the

possible parings that we tested and three of their basic key attributes.

B. Results

Applying SMA to the 273 datasets at a resolution threshold value of $\tau_d = .05$ for all dimensions d using the exponential distance function $e^{-\Delta^r_{[d]}(\overrightarrow{x_j},\overrightarrow{x_k})}$, SMA identified the three most diagnostic combination of dimensions for each dataset. We compared these results to the three best dimensions as selected by the Exhaustive, PCA, NLPCA, KPCA, and mRMR methods. The performance of the 30 possible pairs (6x5=30) of dimensionality-reduction and classification techniques, including SMA, were compared in terms of the number of classification errors that each pair yielded on the three dimensions identified by each reduction method. Because the supervised classifications methods required training sets, we used the following schema to select these. First, 70% of the data points, evenly distributed among the two subcategories of the data set in question, were used for training, and the remaining 30% were used for testing in classification tasks. Similarly, the clustering tasks were performed using 30% of the data points evenly distributed between both categories. This sub-sampling procedure was repeated 50 times for each dataset and the resulting error rates were averaged.

Each dimensionality reduction method was rated on the basis of the percentage of times that it yielded a higher or equal (within .05%) error rate in its best pairing possible when compared to all the 30 pairings. These results are shown in figures 2 and 3, where Fig. 2 illustrates the performance measure on the "actual" or original sets and Fig. 3 illustrates the performance measure with respect to the simulated sets. Note that figures 2 and 3 show only the results of the pairing of dimensionality reduction and classification/clustering technique that combined gives the highest accuracy. The actual values are shown in Table V, for actual datasets, and Table VI, for simulated datasets.

From these graphs, it is clear that the most effective SMA pairing outperformed nearly all possible pairings of reduction and classification techniques. Most notable among these were pairings involving KPCA, NLPCA, and PCA across all three types of actual datasets as illustrated in Fig. 2. On average, the best performing combinations were the paring of SMA and LDA (1.1%) for all three actual datasets.

On the simulated datasets, SMA outperformed or equaled all thirty pairings when applied to CVRD and Alzheimer's data. On average, the best performing combinations were the parings of SMA and LDA (.088%) and SMA with SVM (.087%). On the simulated cancer datasets (WDBC), SMA only outperformed KPCA and NLPCA. However, as can be seen in Fig. 3, the performance differences between the other two methods (PCA and mRMR) on the simulated cancer datasets drawn from different distributions were small. Closer inspection of Table VII shows that SMA does about as well as mRMR on the simulated cancer data as well. In short, SMA coupled with a suitable classification method, outperforms every dimensionality reduction and classification pairing as tested on each of the actual datasets. Moreover, it outperforms or equals, on average, nearly every pairing with respect to all the simulated datasets.

TABLE IV. TESTED REDUCTION AND CLASSIFICATION METHODS

Reduction Method	Strongly-Supervised	Weakly-Supervised	Parametric
Reduction Method			
Exhaustive[a]	NO	NO	NO
PCA	NO	NO	NO
Non-linear PCA	NO	NO	NO
Kernel PCA	NO	NO	NO
mRMR	NO	YES	NO
SMA	NO	NO*	NO
Classification Method			
LDA	YES	YES	YES
KNN	YES	YES	NO
SVM	YES	YES	YES
K-means	NO	NO	NO
Hierarchical	NO	NO	NO

[a] The term "exhaustive" refers to the brute force process of testing every possible combination of three dimensions from the total set of dimensions. *Dependent on chosen heuristic (the heuristic described in this paper involves weak supervision because the datasets are partitioned into two classes; however, similar to identical results are achievable without partitioning the datasets using alternative heuristics).

VII. CONCLUSION

We utilized GIST, a well-studied and accurate theory of human classification behavior, to analyze multivariate datasets. With the addition of a simple heuristic, we called the approach SMA. We then compared SMA to well-known and state of the art non-parametric approaches to feature selection. In turn, their performance was tested using the following classification and clustering methods: linear discriminant analysis, k-nearest neighbor, support vector machine, K-means and hierarchical clustering. In addition to being parsimonious and intuitive, the overall performance of SMA was as good as or better than the aforementioned combined methods. As an added bonus, SMA provides an integrated measure of data complexity. Finally, SMA is non-parametric and non-probabilistic and, therefore, not susceptible to the limitations of either condition.

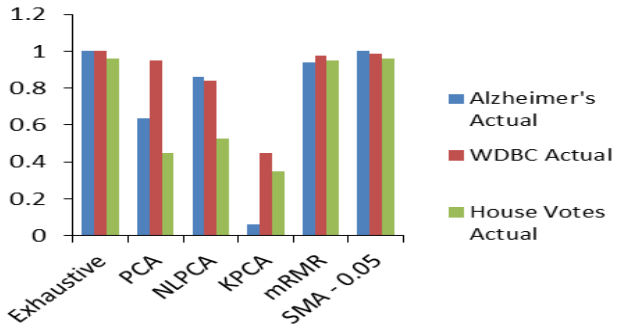

Fig. 2. Summary of Performance for each Reduction Method with Respect to the Average Accuracy Rates of All Three Actual Datasets Tested.

TABLE V. PERFORMANCE ACCURACY ON ACTUAL DATASETS

Dataset	Exhaustive	PCA	NL PCA	K PCA	m RMR	SMA
Alzheimer's	1	0.638	0.863	0.063	0.938	1
WDBC	1	0.95	0.838	0.45	0.975	0.988
CVRD	0.963	0.45	0.525	0.35	0.95	0.963
Average	0.988	0.679	0.742	0.288	0.954	0.983

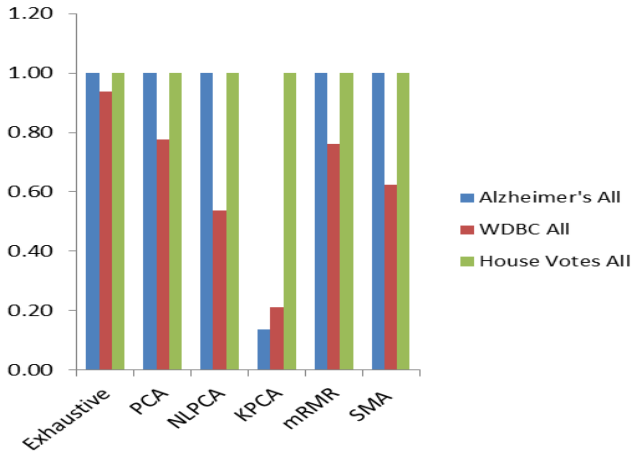

Fig. 3. Summary of Performance for each Reduction Method with Respect to the Average Accuracy Rates of All the Simulated Datasets Tested.

TABLE VI. PERFORMANCE ACCURACY ON SIMULATED DATASETS

Dataset	Exhaustive	PCA	NL PCA	K PCA	m RMR	SMA
Alzheimer's	1	1	1	0.138	1	1
WDBC	0.938	0.775	0.538	0.213	0.763	0.625
CVRD	1	1	1	1	1	1
Average	0.979	0.925	0.846	0.45	0.921	0.858

TABLE VII. PERFORMANCE ERROR ON WDBC DATASET

	LDA	KNN	SVM	K-means	Hierarchical
Exhaustive	1.11	2.32	0.74	3.68	8.88
PCA	2.74	2.92	2.61	6.34	10.37
NLPCA	3.06	6.42	4.97	6.31	9.90
KPCA	14.80	3.34	14.97	14.87	14.88
mRMR	1.88	6.46	1.40	6.37	9.83
SMA	2.10	6.42	1.66	6.35	9.80

ACKNOWLEDGMENT

We wish to thank Karina-Mikayla Barcus and Charles Doan for helping with the data analysis in this project.

REFERENCES

[1] R. Vigo, "Categorical invariance and structural complexity in human concept learning", *Journal of Mathematical Psychology*, 53 (4) 203–221, 2009.

[2] R. Vigo, "Towards a law of invariance in human conceptual behavior", In L. Carlson, C. Hölscher, T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, Austin, TX: Cognitive Science Society, 2580-2585, 2011.

[3] R. Vigo, "Representational information: a new general notion and measure of information", *Information Sciences*, 181, 4847-4859, 2011.

[4] R. Vigo, "Complexity over Uncertainty in Generalized Representational Information Theory (GRIT): A Structure-Sensitive General Theory of Information". *Information*, 4, 1-30, 2012.

[5] R. Vigo, "The GIST (Generalized Invariance Structure Theory) of Concepts", *Cognition*, 129(1), 138-162, 2013.

[6] R. Vigo, "Mathematical Principles of Human Conceptual Behavior: The Structural Nature of Conceptual Representation and Processing", *Scientific Psychology Series*; Routledge, Taylor and Francis, New York and London, 2014.

[7] I. T. Jolliffe, *Principal Component Analysis*. New York: Springer, 2002.

[8] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, no. 2, pp. 233-243, 1991.

[9] B. Schölkopf and A. J. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press, 2002.

[10] O. Alter, P. O. Brown and D. Botstein, "Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling," *Proc. Natl. Acad. Sci.*, vol. 97, no. 18, pp. 10101–10106, 2000.

[11] H. Peng, F. Long and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.27, no.8, pp.1226-1238, Aug. 2005

[12] R. N. Shepard, A. K. Romney, and S. B. Nerlove, Eds., *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences. Vol. I: Theory*. New York: Seminar Press, 1972.

[13] T. J. Hastie, R. J. Tibshirani and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2009.

[14] C. D. Manning, and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT press, 1999.

[15] UCI Learning Repository, http://www.ics.uci.edu/mlearn/MLSummary.html, 2015.

[16] The Comprehensive R Archive Network, http://cran.r-project.org, 2015.

Errata:

1. Capital Phi in Equations 17 and 18 stating Vigo's invariance law was not squared in the published paper. Here these equations have been corrected.

2. The entry in Table IV corresponding to SMA has been changed to reflect the fact that SMA can be either weakly-supervised or not supervised at all with similar results. The following has been added to the caption: "*Dependent on chosen heuristic (the heuristic described in this paper involves weak supervision because the datasets are partitioned into two classes; however, similar to identical results are achievable without partitioning the datasets) using alternative heuristics."

3. Citations 1 to 4 are to articles containing theoretical work that led to the development of GIST; they are not articles on GIST.