LLNL-CONF-682240

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Finding Optimal Set of Axis-Aligned Projections for High-Dimensional Data

J. J. Thiagarajan, K. N. Ramamurthy, P. T. Bremer

February 8, 2016

Eurovis 2016
Groningen, Netherlands
June 6, 2016 through June 10, 2016

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

EUROGRAPHICS 2016 / K-L. Ma, G. Santucci, and J. J. van Wijk
(Guest Editors)

Volume 35 (2016), Number 3

Finding Optimal Set of Axis-Aligned Projections for High-Dimensional Data

Submission ID: 334

Abstract

Visualizing key aspects of high-dimensional data remains one of the most intuitive and popular approaches to gain new insights. In particular, creating a small set of two-dimensional, linear projections has shown to be effective. Existing techniques either select good projections from a large subset of possible projections, which can result in redundant views and do not scale to a large number of dimensions; or optimize for diverse views without considering potential distortions. Furthermore, most recent techniques focus on creating linear projections which can be difficult to interpret by application scientists.

We propose a two-stage, data driven approach to generate optimal sets of 2D axis-aligned projections. It can scale well to higher dimensions and produce a compact, yet diverse, set of high quality projections. We show that despite the restriction to axis-aligned projections our technique recovers the structure of high-dimensional data well, while being easily interpretable. We illustrate the power of our technique using several high-dimensional datasets.

1. Introduction

Exploring and analyzing high dimensional, multi-variate data has become a ubiquitous problem in science and engineering. As the dimension grows, many traditional analysis approaches either become intractable or their results difficult to interpret. Instead, techniques from information visualization are used to allow users to explore various aspects of their data. In particular, 2D linear projections, i.e. scatterplots, are widely used as they are comparatively easy to create and more intuitive than, for example, advanced non-linear embeddings. However, even exploring all axis-aligned projections becomes infeasible as the dimension grows and the space of linear projections is infinitely larger. Furthermore, many of the projections are redundant, producing virtually the same or rigid transformations of a scatterplot, and typically a large portion heavily distorts the data through unacceptable projection artifacts. Since ideally, only a very limited number of projections should be presented to users, selecting an informative set of projections is crucial.

One popular set of approaches is based on ranking projections based on various quality measures designed to find correlations, clusters, or similar features [FT74a, WAG05, WAG06, Guo03]. In general, such techniques create a large set of random projections and select the k highest ranked ones to present to the user. However, as the dimension grows the candidate set required for an adequate sampling of the space of projections becomes intractably large. Furthermore, some of the resulting projections may be redundant, while others with complementary information but lower ranking are ignored. Lehmann and Theisel [LT16] address the latter problem by iteratively selecting projections as diverse from an existing set as possible. More specifically, they consider a new projection to be informative if it cannot be obtained by a linear combination of affine transformations of the existing projections. While this technique creates a small set of diverse projections they do not take the quality of the projection into account. As such it is possible to create a diverse set of significantly distorted views. Finally, many approaches consider linear projections even though scatterplots of an arbitrary linear projection are often difficult to comprehend for users. For example, in the physical sciences users are primarily looking for relationships between attributes, e.g. the temperature, pressure, etc. and data embedded into non-intuitive subspaces is unproductive at best and confusing at worst. While ranking based approach can easily be restricted to axis-aligned subspaces the problem of an unmanageable search space and redundant results remains.

To address the challenges discussed above, we propose a two-stage, data-driven approach to find an optimal set of axis-aligned projections. In the first stage, we identify a small set of linear *candidate projections* considering both the quality of a projection, i.e. how well it preserves neighborhoods, and its novelty with respect to previously inferred projections. In the second stage, we exploit the fact that each

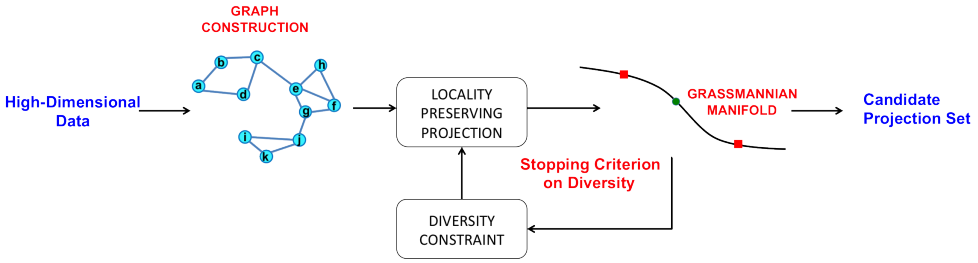


Figure 1: *Stage 1 - Finding a set of candidate linear projections using the proposed optimization strategy.*

linear projection is equivalent to a basis for the corresponding 2D subspace, and represent each projection as a sparse linear combination of axis-aligned subspaces of the high-dimensional data. Finally, we adopt tools from Dempster-Schafer theory [S*76, LHMBVM97] to define a measure of *evidence* for each axis-aligned subspace that indicates its impact in representing the candidate linear subspaces of the data. Our contributions in detail are:

- A mathematical framework to find a diverse set of linear projections by simultaneously maximizing the quality of the projection and the distance from previously selected ones;
- A greedy algorithm to represent a linear subspace as a sparse set of axis-aligned subspaces. In particular, we develop an Orthogonal Matching Pursuit (OMP) technique in the space of all linear subspaces;
- A novel approach that can reveal axis-aligned subspaces that have slightly weaker contributions to the representation of linear subspaces, but are nevertheless important in visualizing the data;
- A new *evidence* measure used to sub-select and rank axis-aligned subspaces by encoding its importance in describing the candidate projection.

We evaluate the proposed approach using a number of standard high-dimensional datasets and present two case studies using a protein homology dataset, and a simulation ensemble from the National Ignition Campaign.

2. Related Work

Given the challenges in directly visualizing high-dimensional data, techniques that identify low dimensional projections are a logical next step. A variety of dimension reduction techniques have been proposed in the data mining and machine learning literature, broadly classified into linear projection [Fuk90] and non-linear projection methods [SL00, BN01]. However, in many cases the intrinsic dimension of the data is higher than two in which case even the best projection will not necessarily reveal the true structure of the data. Hence, approaches that resort to creating multiple 2D projections, in lieu of a single reduced dimensional representation, have become popular.

Optimized 2-Dimensional Projections: Several 2D projections can be strung together into a sequence using interpolation methods providing the user with an overview of high-dimensional data. These interpolation methods are called tours [CBCH95]. In the *Grand Tour*, we randomly navigate through 2D projections of the data, and hence it is possible that we miss the important aspects [Asi85]. In the *projection pursuit guided tour*, we seek for 2D projections that expose the interesting features of the data by optimizing a criterion function [FT74b]. One example is the *Holes* criterion which finds a projection where there is a gap between two clusters of points. However, both these approaches require time-consuming exploration by the user, and the time spent grows exponentially with the number of data dimensions. Recently, in [LT16], the authors propose to provide an optimized tour of the data by identifying the low-dimensional projections that are not redundant. In particular, they identify 2D views of the data that are most dissimilar after accounting for the rotation and translation of the views.

In contrast to these approaches, we propose to provide a set of axis-aligned subspaces that are inherently easier to interpret compared to arbitrary linear projections. Further, we navigate the manifold of linear subspaces of the data in a principled manner using Grassmann manifold theory. Our criteria are comprised of both the quality of projection and maximal separation from the previously selected ones.

Analyzing Axis-Aligned Subspaces: Another class of high-dimensional analysis techniques is based on the assumption that for any one feature only a small subset of attributes might be relevant. These methods typically examine all (or a large number of) possible subsets to find the most interesting ones. Methods such as CLIQUE [AGGR98], SUBCLU [KKK04], PROCLUS [AWY*99], or SURFING [BPR*04] first cluster the subsets by similarity of the data projected into the subspace [KKZ09]. They then choose a representative for each cluster based on various filtering criteria. In high dimensions, the number of possible subspaces can be large and hence it is common to select a small set of them that are structurally "interesting". Some commonly adopted criteria include density variation of samples across regular partitions [CFZ99] or distribution of k-nearest neighbor distances [BPR*04]. Tatu et al. [TMF*12] pro-

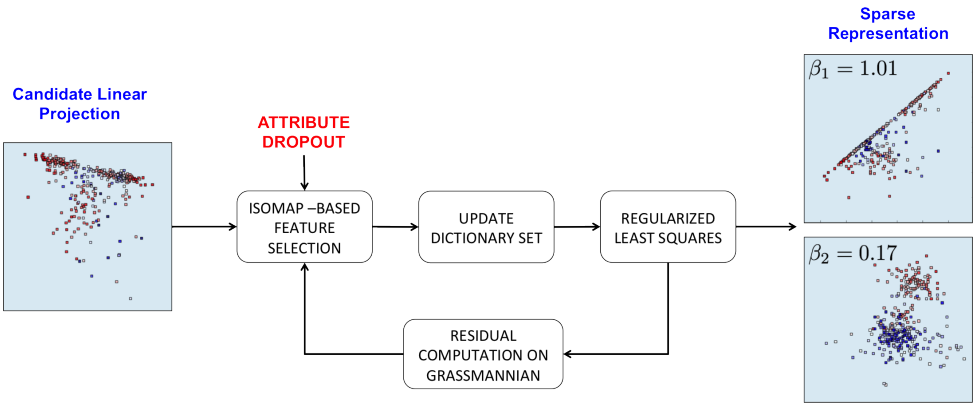


Figure 2: *Stage 2 -Determining the set of axis-aligned subspaces that approximates a given candidate projection using the proposed Grassmannian Sparse Representation algorithm.*

posed a novel visualization approach that uses the SURFING algorithm to find interesting subspaces, and then allows users to navigate through summaries of these subspaces to detect clusters. In [YRWG13], the authors augmented this method with dimension exploration to identify correlation structures between data as well as between dimensions.

However, while axis-aligned subspaces are comparatively easy to interpret, the metrics used to rank them can be misleading because of the projection artifacts, and these methods do not scale well with increasing data dimensions. Instead of using quality metrics to measure how well a projection preserves a global structure, we use principles from Dempster-Shafer evidence theory [S*76] to identify the significance of axis-aligned subspaces in approximating any candidate linear projection.

3. Overview

In this paper, our goal is to find a set of 2D axis-aligned projections, i.e., scatter plot with respect to the original data attributes, to visualize high-dimensional data. We develop a two stage approach, that first computes projections on a diverse set of 2D linear subspaces that can preserve the neighborhood topology of data (Figure 1), and then determines a small set of axis-aligned subspaces that can approximate each of these linear subspaces (Figure 2). Since all 2D linear subspaces are points on the *Grassmann manifold* (or *Grassmannian*), we can develop a principled mathematical formulation for finding a diverse set of candidate subspaces by navigating the Grassmannian. The second stage uses sparse approximation theory to identify a small set of axis-aligned subspaces to approximate the candidates. Finally, we adopt ideas from the Dempster-Shafer theory (DST) [S*76, LHMBVM97] to compute a belief measure per axis-aligned subspace that indicates how well that subspace approximates any of the candidate subspaces. By design, some strong attributes may dominate in the approximation of linear subspaces, thereby drowning out the other attributes that may be important for analysis. To overcome this problem, we propose to employ a dropout scheme, a recently popular regularization strategy in machine learning [SHK*14], during the estimation of axis-aligned subspaces.

Clarifying the terminology used in this paper, by the term subspaces, we refer to the actual data subspaces which are denoted using their generating orthonormal basis. For example, the matrix with orthonormal columns, $\mathbf{V} \in \mathbb{R}^{d\times 2}$ is the generating basis for the 2D subspace spanned by it in the d-dimensional space. The projection of any data point $\mathbf{x} \in \mathbb{R}^d$ onto $\mathbf{V}$ is given by $\mathbf{y} = \mathbf{V}^T\mathbf{x}$.

4. Optimizing Candidate Linear Subspaces on the Grassmann Manifold

Let us denote a high-dimensional (HD) dataset of T samples in $d-$dimensions by the matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$. Our criteria in obtaining candidate linear projections is that the local neighborhood of the data points is preserved in the 2D visualization after the projection. The problem of obtaining projections that can preserve the local neighborhood structure in the data has been studied extensively in the machine learning literature. Some approaches that can be used to achieve this are the locality preserving projections (LPP) [Niy04] and neighborhood preserving embeddings (NPE) [HCYZ05]. These methods require the local neighborhood is encoded as an affinity graph G with the data points in $\mathbf{X}$ as its nodes. The graph is represented using the symmetric adjacency matrix $\mathbf{W} \in \mathbb{R}^{T\times T}$, which is typically sparse. The optimal projection can be obtained using the eigen spectrum of the graph Laplacian matrix [He05]. For example, in LPP, the Laplacian matrix is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is the diagonal degree matrix of the graph with $D_{ii} = \sum_{i\neq j} W_{ij}$. Denoting the low-dimensional

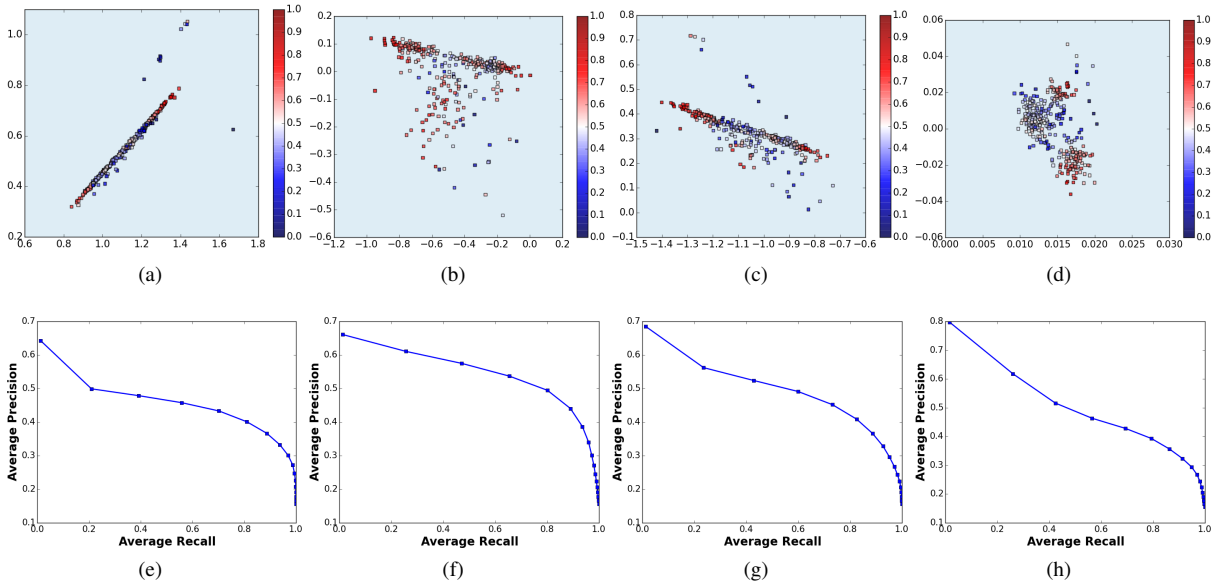


Figure 3: *UCI Ecoli dataset - Diverse set of linear projections identified using the proposed Grassmann optimization technique. In addition to being diverse, we also notice that each of them preserves the neighborhood structure effectively, as evidenced by the Average Recall vs Average Precision curves.*

projection of the HD data as $\mathbf{Y} = \mathbf{V}^T\mathbf{X}$, the generating basis $\mathbf{V} \in \mathbb{R}^{d\times 2}$ is optimized as

$$\mathbf{V} = \underset{\mathbf{V}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{V}=\mathbf{I}}{\arg\min} \ \mathrm{Tr}(\mathbf{V}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{V}). \tag{1}$$

The orthonormality of $\mathbf{V}$ is ensured by enforcing the additional constraint $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. Note that, in the ideal scenario, the local neighborhood of $\mathbf{Y}$ and $\mathbf{X}$ will look the same.

However, since the intrinsic dimensionality of data is often greater than 2, any 2D projection will invariably result in information loss. Hence, it is necessary to consider multiple 2D projections to obtain a more comprehensive view of the data. Similar to the approach in [LT16], we incorporate diversity (or dissimilarity) as a quality measure to infer multiple projections. However, instead of comparing two 2D projections, we propose to directly compare their corresponding subspaces on the Grassmannian manifold [YL14]. The Grassmannian manifold $\mathrm{Gr}(2,d)$ parameterizes all 2D subspaces of $\mathbb{R}^d$ with a well-defined geodesic distance associated to it. Since each projection corresponds to a unique point on the Grassmannian, we can sequentially infer a diverse set of projections that are maximally separated with respect to the Grassmannian distance. We use a particular version of this distance - the squared chordal distance defined as

$$\begin{aligned}\rho^2(\mathbf{V}_j,\mathbf{V}_i) &= d - \|\mathbf{V}_i^T\mathbf{V}_j\|_F^2,\\ &= d - \mathrm{Tr}\left(\mathbf{V}_j^T\mathbf{V}_i\mathbf{V}_i^T\mathbf{V}_j\right).\end{aligned}$$

Here $\mathbf{V}_i$ and $\mathbf{V}_j$ are two points in $\mathrm{Gr}(d,2)$ and equivalently the orthonormal bases of two 2D subspaces.

As described earlier, we sequentially infer a set of diverse candidate projections - first using (1) to obtain a neighborhood-preserving 2D projection for $\mathbf{X}$, and then finding a subspace that provides not only a good fidelity for the data but also is far from the subspaces computed so far. Assuming that we need to compute the $(j+1)^{\text{th}}$ subspace, its diversity is measured as the sum of distances between that subspace and all the previous $i = \{1,\ldots,j\}$ subspaces, i.e.,

$$\sum_{i=1}^{j}\rho^2(\mathbf{V}_{j+1},\mathbf{V}_i) = jd - \mathrm{Tr}\left(\mathbf{V}_{j+1}^T\sum_{i=1}^{j}\left(\mathbf{V}_i\mathbf{V}_i^T\right)\mathbf{V}_{j+1}\right).$$

Hence, the optimization problem for computing the $j+1^{\text{th}}$ subspace is

$$\mathbf{V}_{j+1} = \underset{\mathbf{V}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{V}=\mathbf{I}}{\arg\min} \ \mathrm{Tr}\left(\mathbf{V}^T\left(\mathbf{X}\mathbf{L}\mathbf{X}^T + \alpha\sum_{i=1}^{j}\left(\mathbf{V}_i\mathbf{V}_i^T\right)\right)\mathbf{V}\right), \tag{2}$$

where α is the trade-off parameter between the data representation fidelity and the dissimilarity between subspaces. This can be solved efficiently using trace ratio optimization approaches such as the Decomposed Newton Method [JNZ09].

For demonstration, let us consider the UCI Ecoli dataset [Lic13] of protein localization sites, which contains 336 samples in 7 dimensions. We used the optimization in (2) to infer a diverse set of candidate projections, and the results

are shown in Figure 3. In addition to showing the projections, we include the Average Recall vs Average Precision curves of the neighborhood preserved in each of the projections (details can be found in 5.4).

5. Optimal Set of Axis-Aligned Subspaces

Though the linear projections obtained in the previous section are effective in preserving the local neighborhood of the data, they are not easily interpretable, since each axis in the 2D scatter plot is a linear combination of the original attributes. While considering projections directly defined using the original data attributes can be natural solution to ensure interpretability, exploring the space of all axis-aligned subspaces is very challenging. Hence, we propose to represent each candidate linear subspace using a sparse subset of axis-aligned subspaces. The sparse subset is estimated using the greedy Orthogonal Matching Pursuit (OMP) algorithm on the Grassmannian (Figure 2) inspired by the sparse approximation approach proposed in [HSSL13]. The greedy algorithm avoids the exhaustive search on all possible axis-aligned subspaces, by computing one axis-aligned subspace at a time to best represent the linear subspace. This step can be solved as a feature selection problem. Following this, we evaluate the coefficient value for that subspace using regularized least squares, and compute the residual, which can be projected back to the Grassmannian. Another axis-aligned subspace is chosen using the same procedure to approximate the residual and this process continues until a stopping criteria on the number of subspaces or the approximation error is met.

5.1. Sparse Approximation on Grassmann Manifold

Since a single axis-aligned subspace is not sufficient to approximate the linear subspace represented by $\mathbf{V}$, we compute a sparse representation in terms of the axis-aligned subspaces. We denote the set of all 2D axis-aligned subspaces by $\{\mathbf{Z}_i\}_{i\in\mathcal{Q}}$, where $\mathcal{Q}$ is the index set of size $\binom{d}{2}$. The sparse approximation on the Grassmannian is

$$\beta = \underset{\{\beta_i\}}{\arg\min} \|\mathbf{V}\mathbf{V}^T - \sum_{i=1}^{|\mathcal{Q}|} \beta_i \mathbf{Z}_i \mathbf{Z}_i^T\|_F^2, \quad \text{subject to } \|\beta\|_0 \leq L. \tag{3}$$

Here $\beta \in \mathbb{R}^{|\mathcal{Q}|}$ is the sparse coefficient vector, $\|.\|_0$ denotes the ℓ_0 norm that counts the total number of non-zero entries in a vector and L is the maximum number of axis-aligned subspaces that can be chosen (stopping criterion). The reason for using a greedy algorithm to solve (3) instead of the ℓ_1 relaxation used in [HSSL13] is that the ℓ_1 minimization will be computationally expensive since $|\mathcal{Q}| = \binom{d}{2}$. The full algorithm is provided in Algorithm 5.1

Algorithm 5.1 Optimize for β given $\mathbf{V}$ and L

1. Initialize $\Omega = \varnothing$, $\mathbf{U} = \mathbf{V}$

Greedy update of β and Ω:

2. While **Stopping Criterion** is not met:
   a. Identify index $i_{\hat{\alpha}}$ of $\mathbf{Z}_{i_{\hat{\alpha}}}$ to approximate $\mathbf{U}$ using (6)
   b. Update non-zero support set: $\Omega \leftarrow \Omega \cup i_{\hat{\alpha}}$
   c. Compute β_Ω using (7)
   d. Residual computation: $\mathbf{R} = \mathbf{V}\mathbf{V}^T - \sum_{i\in\Omega} \beta_i \mathbf{Z}_i \mathbf{Z}_i^T$
   e. Reprojection: Update $\mathbf{U}$ with top two principal eigen vectors of $\mathbf{R}$
3. Output Ω and β_Ω

Choosing a single axis-aligned subspace: We denote the linear subspace to be represented using the axis aligned subspaces as $\mathbf{U}$. Interestingly, the problem of choosing an axis-aligned 2D subspace can be viewed as a feature selection problem. In particular, we use the Isomap-based feature selection strategy proposed in [DD14], which attempts to preserve the neighborhood structure in the subspace $\mathbf{U}^T\mathbf{X}$. We begin by defining the pairwise point difference between two data points $\mathbf{x}_i$ and $\mathbf{x}_j$, also known as the *secant* as

$$s_{ij} = \frac{\mathbf{x}_i - \mathbf{x}_j}{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}. \tag{4}$$

Clearly, $\sum_{n=1}^{d} s_{ij}^2(n) = 1$. We will define $\mathcal{S}$ to be the set of all possible secants in the data $\mathbf{X}$. Since we are interested in choosing attributes of $\mathbf{X}$ that best preserve the structure in $\mathbf{U}^T\mathbf{X}$, we define $\mathcal{S}_k$ to contain only the secants with the k nearest neighbors in that subspace:

$$\mathcal{S}_k = \{s_{ij} | i = \{1,\ldots,T\}, j \in \mathcal{N}_k(i)\}. \tag{5}$$

Here $\mathcal{N}_k(i)$ contain the indices of the k nearest neighbors of $\mathbf{y}_i$. We construct a matrix $\mathbf{C} \in \mathbb{R}^{|\mathcal{S}_k|\times d}$ such that each of its rows is a secant from the set $\mathcal{S}_k$, squared elementwise. The expected squared norm of the reduced dimensional secant vectors, where 2 out of d dimensions were chosen uniformly at random, is $2/d$. Hence the optimization problem for selecting the 2 dimensions can be formulated as [DD14]

$$\hat{\alpha} = \underset{\alpha}{\arg\min} \|\mathbf{C}\alpha - \frac{2}{d}\mathbf{1}\|_2^2 \quad \text{subject to } \mathbf{1}^T\alpha = 2, \alpha \in \{0,1\}^d, \tag{6}$$

where α is the masking vector that will have a value of one at positions corresponding to the two selected dimensions. This is a computationally intensive binary integer program which can be solved greedily by selecting one dimension of the secant at a time. The mask vector $\hat{\alpha}$ has a one-to-one correspondence with the index set $\mathcal{Q}$, and let us denote the corresponding index by $i_{\hat{\alpha}}$. The axis-aligned subspace obtained is denoted by $\mathbf{Z}_{i_{\hat{\alpha}}} \in \mathbb{R}^{d\times 2}$.

Updating coefficients and residual: Following Algorithm 5.1, the support set Ω of the chosen axis-aligned subspaces is updated. The coefficient β_Ω is computed using regularized least squares:

$$\hat{\beta}_\Omega = \underset{\beta}{\arg\min} \|\mathbf{V}\mathbf{V}^T - \sum_{i\in\Omega} \beta_i \mathbf{Z}_i \mathbf{Z}_i^T\|_F^2 + \lambda\|\beta_\Omega\|_2. \tag{7}$$

The residual $\mathbf{R}$ is computed and reprojected to the Grassmannian using the steps in Algorithm 5.1. The process of choosing another axis-aligned subspace and updating the coefficients is repeated until a stopping criterion on L or the approximation error is met.

5.2. Inferring Evidence Measures

Dempster-Shafer theory (DST) is a general framework for reasoning with uncertainty [S*76], which we will utilize to understand the degree of belief of each axis-aligned subspaces in approximating the candidate subspaces. Let Θ be the universal set of all hypotheses, i.e., the set of all axis-aligned subspaces in our case, and 2^{Θ} be its power set. A probability mass can be assigned to every hypothesis $A \in 2^{\Theta}$ such that,

$$\mu(\varnothing) = 0, \sum_{A \in 2^{\Theta}} \mu(A) = 1, \tag{8}$$

where $\varnothing$ denotes the empty set. This measure provides the confidence that hypothesis A is true. Using DST, we can compute the uncertainty of the axis-aligned subspaces in representing a linear subspace using the belief (bel(.)) function, which is the confidence on that hypothesis being supported by strong evidence. The belief function is defined as

$$\text{bel}(A) = \sum_{B \subseteq A} \mu(B). \tag{9}$$

Using principles from DST, we can easily combine the evidence from multiple sources. In our case, this corresponds to combining beliefs of an axis-aligned subspace in describing multiple candidate subspaces.

For the chosen axis-aligned subspaces given by Ω of a candidate subspace $\mathbf{V}$, the mass corresponding to the axis-aligned subspace $\mathbf{Z}_i$, where $i \in \Omega$ is given as $\mu(\mathbf{Z}_i)$. This is estimated by computing the reconstruction error using $\mathbf{Z}_i$ as $\mathbf{e}^i = \|\mathbf{V}\mathbf{V}^T - \beta_i \mathbf{Z}_i \mathbf{Z}_i^T\|_F$ and defining the mass to be

$$\eta = \eta_0 \exp(-\gamma \mathbf{e}^i). \tag{10}$$

Here η_0 is a parameter in the interval $[0,1]$ (chosen closer to 1) that upper bounds the mass of any single hypothesis.

We apply Dempster's combination rule to combine beliefs of $\mathbf{Z}_i$ from all candidate linear projections. Assuming that there are P linear subspaces denoted by the orthonormal bases $\{\mathbf{V}_i\}_{i=1}^P$, the total mass can be accumulated as

$$\mu(\mathbf{Z}_i) = 1 - \prod_{j=1}^{P} (1 - \eta_0 \exp(-\gamma \mathbf{e}_j^i)). \tag{11}$$

Here $\mathbf{e}_j^i$ denotes the reconstruction error obtained for the jth candidate projection using the ith axis-aligned subspace. The normalized evidence measure of the axis-aligned subspace $\mathbf{Z}_i$ is given by

$$evid[i] = \frac{\mu(\mathbf{Z}_i)}{\max\left(\{\mu(\mathbf{Z}_j)\}_{j=1}^{|\mathcal{Q}|}\right)}, \forall i = 1, \cdots, |\mathcal{Q}| \tag{12}$$

5.3. Random Dropout Regularization

By design, our sparse approximation approach will not choose two different axis-aligned subspaces with shared variables if the resulting structures are very similar. However, from an anyalyst's viewpoint, it is important to have both an compact set of projections to look at, and the ability to easily mine for projections with similar structures. To this end, we develop a dropout a scheme where a small percentage of feature dimensions are randomly dropped when computing the sparse representations. Random dropout schemes are commonly used to improve the robustness of machine learning algorithms such as random forests [Bre01] and deep neural networks [SHK*14]. We repeat this process multiple times, $R = 100$, by choosing a different random subset of attributes to drop, and generate an ensemble of sparse representations. In all experiments reported in this paper, we used a dropout of 10% and computed the accumulated beliefs for each of the runs independently using (11). Using the ensemble, we infer the overall belief for each hypothesis using a max-pooling strategy as follows:

$$\mu(\tilde{\mathbf{Z}}_i) = \max\left(\{\mu(\mathbf{Z}_i^r)\}_{r=1}^{R}\right). \tag{13}$$

This makes the uncertainty quantification using DST more robust by providing a number of alternate beliefs for each hypothesis (axis-aligned subspace).The final evidence measures are obtained using the normalization in (12).

5.4. Augmenting scatterplot visualization with precision-recall

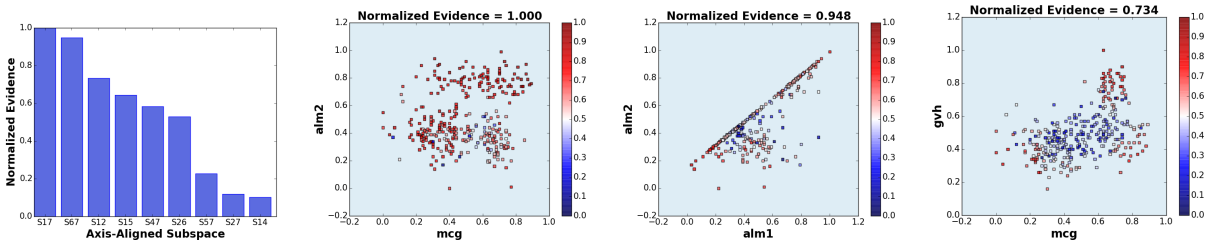
The fidelity of 2D projections can be effectively visualized using a variety of quality measures. In this paper, we consider the use of precision and recall of the preserved neighborhood for each data point. We define the neighborhoods for a point i in the original data and the visualization domain as, $\mathcal{N}_k(i)$, and $\mathcal{N}'_{k'}(i)$ respectively where k and k' denote the number of points in the neighborhood. The precision and recall values are then computed as

$$\text{Precision} = \frac{|\mathcal{N}_k(i) \cap \mathcal{N}'_{k'}(i)|}{k'} \tag{14}$$

$$\text{Recall} = \frac{|\mathcal{N}_k(i) \cap \mathcal{N}'_{k'}(i)|}{k} \tag{15}$$

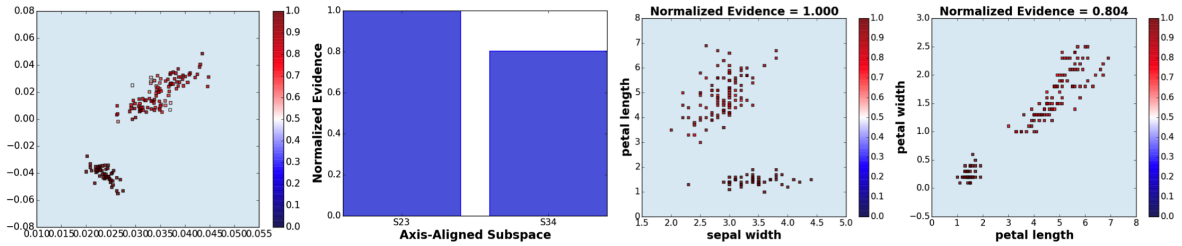
For a given $k\ (= k')$ we estimate fidelity $= 0.5 * \text{Precision} + 0.5 * \text{Recall}$ and color each point in the scatterplot using this measure. We can also generate curves as shown in Figure 3 by showing the average precision/recall values for a fixed k and varying k'. The larger the area under the curve, better the fidelity.

For the UCI Ecoli dataset, we used the proposed approach to infer a compact set of axis-aligned subspaces and evaluated their evidence measures. From the results in Figure 4, we observe that 9 axis-aligned subspaces were picked in total and there is a smooth degradation in the evidence mea-

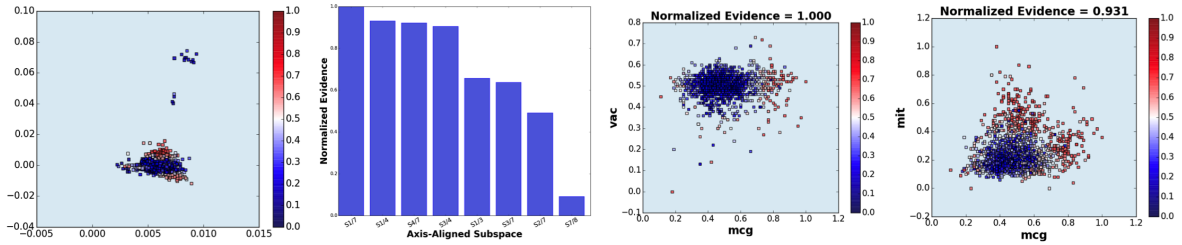


(a) (b) (c) (d)

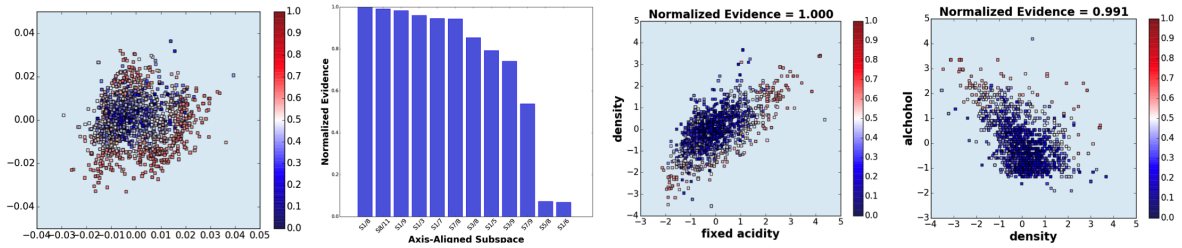
Figure 4: *UCI Ecoli dataset - Optimal set of axis-aligned plots inferred using the proposed algorithm. We show the* 3 *subspaces with the largest normalized evidence scores. Note that, the samples in each of the plots are colored by their fidelity measure.*



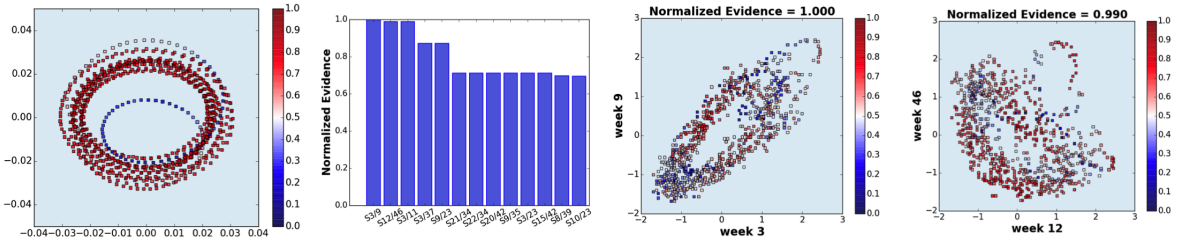
(a) UCI Iris Dataset



(b) UCI Yeast Dataset



(c) UCI Wine Quality Dataset



(d) UCI Seawater Temperature Dataset

Figure 5: *Left-to-right: We show the optimal LPP projection, the set of axis-aligned subspaces chosen by our algorithm along with their evidence measures, and the two axis-aligned scatterplots with the highest evidence.*

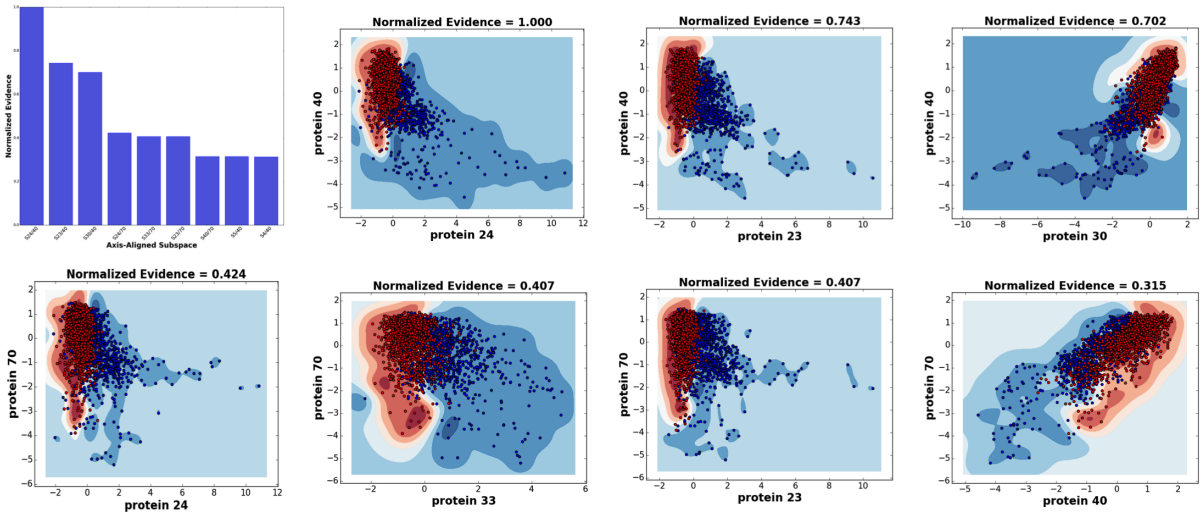


Figure 6: *Protein Homology Dataset Analysis - The first seven axis-aligned subspaces inferred using the proposed algorithm, along with their evidence measure. Using the ground truth labeled data, we train separate RBF SVM classifier in each axis-aligned subspace and show the resulting decision surfaces. Especially for the first couple projections, the classifier is able to separate the classes well, indicating that the projections are well chosen and preserve the main feature of the data.*

sure. Since the scatterplots shown to the user are in terms of the original attributes they are easily interpretable.

6. Experiments

6.1. Standard High-dimensional Datasets

We tested our algorithm on 5 well-known high-dimensional datasets with different dimensions and sample sizes, and the results are reported in Figure 4 and 5. We first extract the diverse set of candidate subspaces that best describe the data. The bar charts show the evidence values for all axis-aligned projections with non-zero evidence for their respective subspace. The first observation is that, in all cases, our method selects a very small number of axis-aligned subspaces and their evidence measures degrade smoothly. If even fewer projections are required the k with highest evidence can be chosen. However, the evidence as computed here is a relative measure for each linear subspace, so global cut-offs would have to be chosen with care. In Figure 5, we also show the optimal linear projection for comparison, obtained using the LPP method described in Section 4. Unless there are some global structures, such as, well-separated clusters, linear projections offer little intuition as the axes are difficult to label and to interpret. On the other hand, analysis with the original data attributes can reveal which attributes lead to the true clusters, and show functional relationships between attributes. We demonstrate this using two case studies.

6.2. Case Study: Protein Homology Dataset

This dataset measures which protein sequences are homologous to a native sequence, evaluated using a variety of scores [pro]. It comprises of 2500 samples out of which 1295 sequences are marked to be homologous, while the rest are not. Note that, we do not use these labels for our analysis but only to validate our results. There are 74 attributes for each sequence which correspond to alignment scores in different regions of the sequence using a variety of strategies. Our approach selects 9 axis-aligned projections, with non-zero evidence values, out of 2701 possible projections. The ability to provide a compact set of scatterplots for comprehensive analysis is crucial when dealing with high-dimensional data. To validate the quality of the chosen projections we fit classifiers (Kernel SVM) in the 2D, projected spaces to the ground-truth labels. As shown in Figure 6, even though each classifier is restricted to only two attributes they are still able to separate the classes well. This indicates that the selected projections indeed preserve the structure of the data.

Table 1: *High-Dimensional datasets used for evaluating the proposed algorithm.*

Dataset	# Dimensions	# Samples
Iris	4	150
Ecoli	7	336
Yeast	8	1484
Wine Quality	13	1599
Seawater	52	823

6.3. Case Study: NIC Engineering Simulation Dataset

The National Ignition Campaign (NIC), a collaboration between Lawrence Livermore, Los Alamos, and Sandia Na-

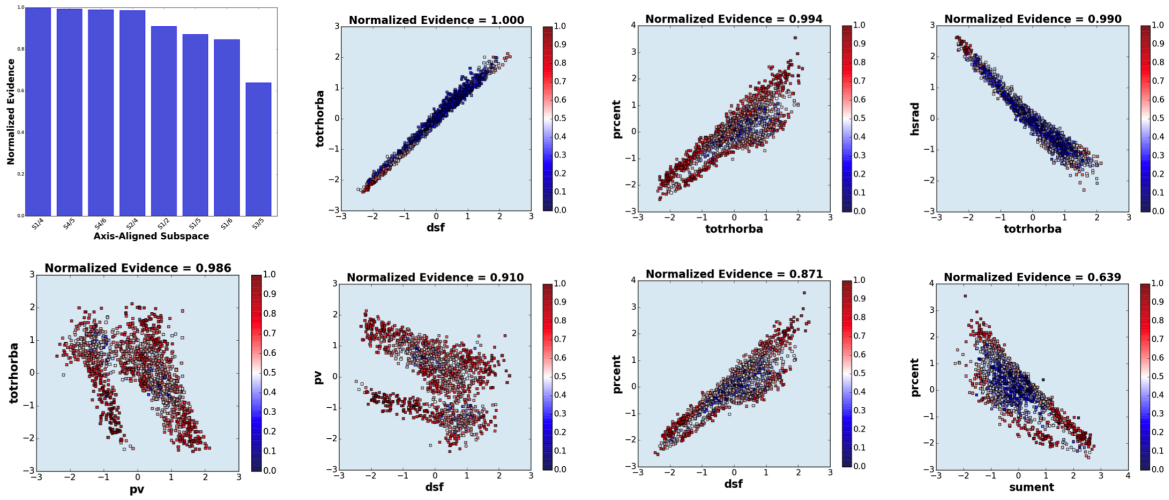


Figure 7: *NIC Simulation Data Analysis - Our proposed approach reveals the presence of two true clusters in the attribute pv and a strong correlation between the attributes dsf and totrhorba, though the projection artifacts mask the neighborhood structure.*

tional Laboratories as well as The University of Rochester and General Atomics, is aimed at demonstrating inertial confinement fusion (ICF), that is, thermonuclear ignition and energy gain in a laboratory setting. Fundamentally, the goal of NIF is to search the parameter space to find the region that leads to near-optimal performance. The dataset considered here is a so called engineering or macro-physics simulation ensemble in which an implosion is simulated using various different in parameters, such as, laser power, pulse shape etc.. From these simulations scientists extract a set of *drivers*, physical quantities thought to determine the behavior of the resulting implosion. These drivers are then analyzed with respect to the energy yield to better understand how to optimize future experiments. The dataset, we consider for our analysis, consists of 1304 samples with 6 drivers: (i) down scatter fraction (dsf), (ii) peak velocity (pv), (iii) entropy (sument), (iv) (totrhorba), (v) pressure at the centre (prcent), and (vi) hotspot radius (hsrad) and various quality metrics most notably the resulting yield.

Figure 7 shows the 8 axis-aligned projections, which describe all candidate linear projections inferred using the optimization of Section 4. The projection with the highest evidence is comprised of the attributes {*dsf*, *totrhorba*} and it reveals a strong linear correlation. Surprisingly, the precision-recall based fidelity measures show that this projection contains significant artifacts. However, a closer look at the related projections {*dsf*, *pv*} and {*pv*, *totrhorba*} shows that both of them contain two distinct clusters with very high fidelity, leading us to the inference that the attributes *dsf* and *totrhorba* are indeed very strongly correlated. Geometrically, there are two different regions in the HD space where these two attributes are strongly correlated, and the neighborhood structure is partially lost due to the projection. However, our algorithm is able to correctly determine that this relationship is valid and in fact assigns to it the highest evidence measure. Consulting with the relevant experts at LLNL, we have found that the presence of two true clusters in only certain attributes, and the attribute correlation have been very useful to them in understanding this data. In addition, earlier attempts at this analysis using linear projections were unsuccessful as the physicists had difficulties interpreting the linear axes, which ultimately led to them abandoning the approach.

7. Conclusions

In this paper, we presented a data-driven approach to select relevant axis-aligned subspaces for visualizing high-dimensional data. While finding a compact, yet diverse, set of meaningful 2D projections can be valuable in information visualization, their interpretability can be significantly improved by using projections in terms of the original attributes. To this end, we presented a two-stage approach, which first identifies a diverse set of high-fidelity linear projections, and then picks axis-aligned subspaces that can effectively describe the neighborhood topology in the candidate subspaces. The proposed approach consistently provides a compact set of views, while revealing insights about the true clusters and complex attribute relationships in the HD data. Our greedy algorithm for finding useful subspaces is robust and can scale to very high dimensions. Though linear projections have been commonly adopted in data analysis, it is not straightforward for an analyst to map the observations in the projected space back to the original data attributes. We believe that our approach is a first step towards bridging the gap between automated projection finding techniques and user interpretability.

Prepared by LLNL under Contract DE-AC52-07NA27344.

References

[AGGR98] AGRAWAL R., GEHRKE J., GUNOPULOS D., RAGHAVAN P.: Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of ACM SIGMOD* (1998), pp. 94–105. 2

[Asi85] ASIMOV D.: The grand tour: a tool for viewing multidimensional data. *SIAM journal on scientific and statistical computing 6*, 1 (1985), 128–143. 2

[AWY*99] AGGARWAL C. C., WOLF J. L., YU P. S., PROCOPIUC C., PARK J. S.: Fast algorithms for projected clustering. In *Proceedings of ACM SIGMOD* (1999), pp. 61–72. 2

[BN01] BELKIN M., NIYOGI P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems* (2001), MIT Press, pp. 585–591. 2

[BPR*04] BAUMGARTNER C., PLANT C., RAILING K., KRIEGEL H.-P., KROGER P.: Subspace selection for clustering high-dimensional data. In *Proceedings of IEEE ICMD* (Nov 2004), pp. 11–18. 2

[Bre01] BREIMAN L.: Random forests. *Machine learning 45*, 1 (2001), 5–32. 6

[CBCH95] COOK D., BUJA A., CABRERA J., HURLEY C.: Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics 4*, 3 (1995), 155–172. 2

[CFZ99] CHENG C.-H., FU A. W., ZHANG Y.: Entropy-based subspace clustering for mining numerical data. In *Proceedings of the ACM SIGKDD* (1999), pp. 84–93. 2

[DD14] DADKHAHI H., DUARTE M. F.: Masking schemes for image manifolds. In *Statistical Signal Processing (SSP), 2014 IEEE Workshop on* (2014), IEEE, pp. 252–255. 5

[FT74a] FRIEDMAN J., TUKEY J.: A projection pursuit algorithm for exploratory data analysis. *IEEE TC C-23*, 9 (1974), 881–890. 1

[FT74b] FRIEDMAN J. H., TUKEY J. W.: A projection pursuit algorithm for exploratory data analysis. 2

[Fuk90] FUKUNAGA K.: *Introduction to statistical pattern recognition*. Academic Press, 1990. 2

[Guo03] GUO D.: Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization 2*, 4 (2003), 232–246. 1

[HCYZ05] HE X., CAI D., YAN S., ZHANG H.-J.: Neighborhood preserving embedding. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on* (2005), vol. 2, IEEE, pp. 1208–1213. 3

[He05] HE X.: *Locality Preserving Projections*. PhD thesis, Chicago, IL, USA, 2005. AAI3195015. 3

[HSSL13] HARANDI M., SANDERSON C., SHEN C., LOVELL B. C.: Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution. In *Computer Vision (ICCV), 2013 IEEE International Conference on* (2013), IEEE, pp. 3120–3127. 5

[JNZ09] JIA Y., NIE F., ZHANG C.: Trace ratio problem revisited. *Neural Networks, IEEE Transactions on 20*, 4 (2009), 729–735. 4

[KKK04] KAILING K., KRIEGEL H.-P., KRÖGER P.: Density-connected subspace clustering for high-dimensional data. In *Proceedings of SDM* (2004), pp. 246–257. 2

[KKZ09] KRIEGEL H.-P., KRÖGER P., ZIMEK A.: Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data 3*, 1 (2009), 1:1–1:58. 2

[LHMBVM97] LE HEGARAT-MASCLE S., BLOCH I., VIDAL-MADJAR D.: Application of dempster-shafer evidence theory to unsupervised classification in multisource remote sensing. *Geoscience and Remote Sensing, IEEE Transactions on 35*, 4 (Jul 1997), 1018–1031. doi:10.1109/36.602544. 2, 3

[Lic13] LICHMAN M.: UCI machine learning repository, 2013. URL: http://archive.ics.uci.edu/ml. 4

[LT16] LEHMANN D. J., THEISEL H.: Optimal sets of projections of high-dimensional data. *Visualization and Computer Graphics, IEEE Transactions on 22*, 1 (2016), 609–618. 1, 2, 4

[Niy04] NIYOGI X.: Locality preserving projections. In *Neural information processing systems* (2004), vol. 16, MIT, p. 153. 3

[pro] [online]URL: http://osmot.cs.cornell.edu/kddcup/datasets.html. 8

[S*76] SHAFER G., ET AL.: *A mathematical theory of evidence*, vol. 1. Princeton university press Princeton, 1976. 2, 3, 6

[SHK*14] SRIVASTAVA N., HINTON G., KRIZHEVSKY A., SUTSKEVER I., SALAKHUTDINOV R.: Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research 15*, 1 (2014), 1929–1958. 3, 6

[SL00] SEUNG H. S., LEE D. D.: The manifold ways of perception. *Science 290*, 5500 (2000), 2268–2269. 2

[TMF*12] TATU A., MAASS F., FÄRBER I., BERTINI E., SCHRECK T., SEIDL T., KEIM D. A.: Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *IEEE VAST* (2012), IEEE Computer Society, pp. 63–72. 2

[WAG05] WILKINSON L., ANAND A., GROSSMAN R. L.: Graph-theoretic scagnostics. In *INFOVIS* (2005), vol. 5, p. 21. 1

[WAG06] WILKINSON L., ANAND A., GROSSMAN R.: High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *Visualization and Computer Graphics, IEEE Transactions on 12*, 6 (2006), 1363–1372. 1

[YL14] YE K., LIM L.-H.: Distance between subspaces of different dimensions. *arXiv preprint arXiv:1407.0900* (2014). 4

[YRWG13] YUAN X., REN D., WANG Z., GUO C.: Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2625–2633. 3