Eurographics/ IEEE-VGTC Symposium on Visualization 2009
H.-C. Hege, I. Hotz, and T. Munzner
(Guest Editors)

Volume 28 (2009), Number 3

Selecting good views of high-dimensional data using class consistency

Mike Sips[1] and Boris Neubert[2] and John P. Lewis[3] and Pat Hanrahan[4]

[1] Max Planck Center for Visual Computing Stanford/Saarbruecken, [2] University of Konstanz, [3] Massey University, [4] Stanford University

Abstract

Many visualization techniques involve mapping high-dimensional data spaces to lower-dimensional views. Unfortunately, mapping a high-dimensional data space into a scatterplot involves a loss of information; or, even worse, it can give a misleading picture of valuable structure in higher dimensions. In this paper, we propose class consistency *as a measure of the quality of the mapping. Class consistency enforces the constraint that classes of n–D data are shown clearly in 2–D scatterplots. We propose two quantitative measures of class consistency, one based on the distance to the class's center of gravity, and another based on the entropies of the spatial distributions of classes. We performed an experiment where users choose good views, and show that class consistency has good precision and recall. We also evaluate both consistency measures over a range of data sets and show that these measures are efficient and robust.*

Categories and Subject Descriptors (according to ACM CCS): Data Mining [I.5.3]: Clustering—User Interfaces [H.5.2]: Evaluation/methodology—

1. Introduction

Today's scientific and business applications produce large datasets with increasing complexity and dimensionality. Visual data exploration techniques have proven to be of high value in gaining insight into these large data sets. The aim of visual data exploration is to tightly couple data analysis techniques and interactive visualization methods, and thus combine two powerful information processing systems: the human mind and the modern computer [KSA04].

A major challenge is how to present high dimensional data to the analyst. Many visualization methods involve mapping high dimensional data to lower-dimensional views. Because graphical displays are composed of two spatial coordinates and a limited number of visual variables such as color, texture, etc., the maximum number of dimensions that can be shown in any one view is roughly 3-8 [Ber84]. And since the dimensionality of the data is often quite high – often tens to hundreds of dimensions – the mapping from data space to display space involves a loss of information. The problem is not just partial information however: projected views can also present *misleading* information, since structures that are separated in higher dimensions are often conflated in the 2–D projection. This leads to a major challenge in visualization: How to map from high dimensions to low dimensions in a way that faithfully represents the data? Given a huge collection of possible views, which view represents the data best?

In this paper we propose *class consistency* as a computable measure of the utility of a given view. The basic idea of class consistency is shown in Figure 1. In this figure the original high-dimensional data is represented as a set of two-dimensional points, with red and green representing two classes of data. The low-dimensional views of the data are represented as the marginal 1–D histogram projections along the axes of the scatterplot. In this example, we consider the horizontal projection to be consistent. That is, the red and green points are projected to regions of the display space that are separable. In contrast, the vertical projection is inconsistent: red and green points are mixed together in this projection. We claim that the horizontal projection is better because view and data are consistent.

In this paper, we assume that each point in high dimensional space has been labeled as belonging to some group. Class labels can be automatically assigned using a clustering algorithm. Since 2–D orthogonal projections allow an intuitive interpretation of the data, orthogonal projections

forming 2–D scatterplots are often used as a starting point in exploratory data analysis. In this study, we consider orthogonal projections forming 2–D scatterplots. The set of all possible views is the $n(n-1)/2$ unique scatterplots in a matrix of scatterplots, or SPLOM [Har75]. Since the number of 2–D scatterplots of real world data is much higher than a human analyst can look at, we want the computer to sort-out consistent views which corresponds to choosing the best scatterplots from the matrix of scatterplots. Selecting consistent views for other types of patterns and views is left as future work.

The contributions of this paper are:

- We propose class consistency as criteria for choosing good views to a class structure in n–D. Class consistency characterizes the extent to which the class neighborhood structure in n–D is preserved in a 2–D scatterplot, and thus avoids to label poor views as good views.
- Since human attention is limited to inspect a small number of scatterplots, class consistency used as measure of goodness facilitates an interactive exploration of a class structure; otherwise a human analyst will be drown in the vast set of 2–D scatterplots.
- We introduce and evaluate two methods for calculating class consistency, a distance based and a distribution based technique. Distribution based class consistency is more general, but more expensive to compute.
- We evaluate class consistency over a range of data sets with different dimensionality. We show that the class consistency measures perform well in practice. First user experiments show that people rank consistent views better than inconsistent views.

2. Related Work

Several different approaches have been proposed for selecting good views of high dimensional projections and embeddings. The first major development in this area was projection pursuit. The idea of projection pursuit is to search for a *good* view to high dimensional data [Fri87].

Several criteria have been used to define good views. Tukey and collaborators used clumpiness as a measure of goodness [FT74]. Clumpiness describes the degree to which data points are concentrated locally while at the same time expanded globally in a 2–D embedding. Clumpiness works well when data points are clustered around cluster centers. Given the more general situation in which data points cluster around lines, curves, curved manifolds, etc., clumpiness tends to prefer 2–D embeddings in which large amounts of classes are mixed. The reason can be seen in the fact that mixing classes create regions with arbitrary high local densities. In contrast to clumpiness, class consistency used as measure of goodness assigns high numerical scores to 2–D views in which classes are separated, and low scores in any other situation. A natural question is to use class consistency as a measure of goodness in projection pursuit to find general projections that are consistent with a class structure in higher dimension, which is left for future work.

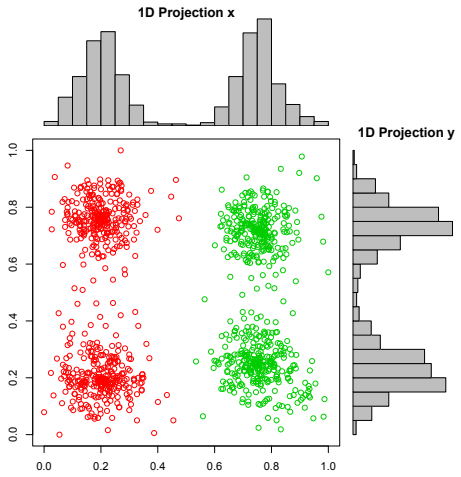
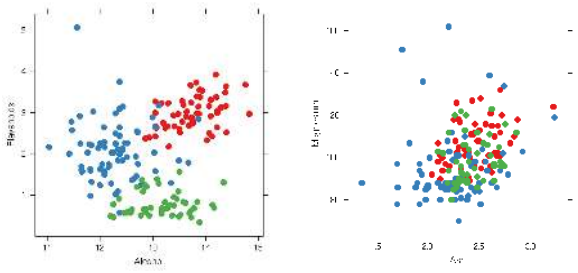


Figure 1: ***Mapping a high-dimensional data space into a low-dimensional space leads in the worst case to a misleading picture of the clusters hidden in the data*** *– the 1–D axis parallel projection of the 2–D cluster model (red and green points) along the x and y axes results in two different views of the data. Although there are two clusters visible in both projections (two peaks in the histograms), only the projection along the x-dimension is consistent with the clusters. The projection in y-direction merges the two clusters and hence is not consistent.*

The grand tour [Asi85] shows an overview of a high-dimensional data space by presenting a sequence of low-dimensional projections. The widely used XGOBI [CBCH95] system combines the grand tour and projection pursuit with a single interactive interface. Although the combination of these two methods is powerful, it is still time consuming to manually explore the space of all 2–D projections in a reasonable amount of time. More fundamentally, there is still disagreement about the measures of goodness used in projection pursuit algorithms. And even if a good measure of goodness is discovered, its value depends on the feasibility to optimizing it.

Scagnostics was proposed by Tukey and Tukey [TT85] as an alternative to projection pursuit. A system using graph-theoretical scagnostic measures was recently described by Wilkinson et al. [WAG06]. In their paper, they compute scagnostics for each scatterplot in a matrix of scatterplots. These graph-theoretic measures are meant to characterize different types of patterns. These measures are then used to create a second matrix of scatterplots. Each scatterplot of the data is represented as a dot in the scatterplots of the scagnostics. The scagnostics approach effectively supports



(a) **DSC**=90 (b) **DSC**=49

Figure 2: ***Consistent (=Good) vs. Non-Consistent (=Poor) View*** *– The data set contains three clusters representing three classes of wine, and 13 attributes describing chemical properties of the wine. The left figure shows the scatterplot for dimensions alcohol and flavanoids. The classes are separated in this view and most data points are located close to class centers, resulting in a consistent view. In the right figure in contrast, in the scatterplot of dimensions ash and magnesium classes are cluttered and not separated, resulting in a poor consistency rating.*

visual data analysis by organizing the different views in the space of patterns. Our approach for detecting class consistency is complementary to scagnostics, and could be used as one of the measures.

The rank-by-feature framework [SS05] helps users to explore 1–D and 2–D orthogonal projections of a high-dimensional data space. It allows the data analyst to examine the 2–D orthogonal projections by ranking these projections according to a criterion chosen by the data analyst. The framework effectively supports the user in exploring valuable correlations between selected dimensions and in finding outliers. Again, our notion of class consistency could easily be added to the rank-by-feature framework.

Cluster-preserving projections seek to generate a single best view of the data by transforming the data into a space that maximizes class separability. Koren and Carmel [KC04] propose an interesting measure of goodness. They weight the distances between points differently depending on whether they have the same label. Given this objective function, they find a linear transformation of the high-dimensional data maximizing inter-cluster and minimizing intra-cluster distances. Their method has a significant advantage in comparison with traditional PCA or MDS, because they capture the cluster structure of the data and additionally the intra-cluster shapes. A similar measure of goodness is proposed in [DMS98]. The authors maximize the distance between the projected means to get a good cluster separation. Given this objective function, they find a 2–D plane parallel to the plane containing the cluster centers. The distances between the cluster centers are persevered under this projection. One problem with general projections and embeddings is that users may have trouble interpreting the display axes (which may be arbitrary linear or nonlinear combinations of the original variables), or the reconstructed 2–D plane (e.g. MDS). In contrast to cluster-preserving projections which seek to generate a single best view, our method scores a set of existing views. Thus, our method can in principle be used as criterion to measure the utility of the transformation for any method that generates a space of possible views.

3. Class Consistency

While many potential criteria that could define a good view are possible, we claim that a good view to a class structure should be at least consistent with that class structure (see Figure 2 for an illustration).

First, we define the data space and the set of views. Let $X \subseteq \mathbf{R}^n$ be a high-dimensional data space consisting of points $x_k = (x_k^1, \cdots, x_k^n)$. Let π_i denote the 1–D orthogonal projection $\mathbf{R}^n \rightarrow \mathbf{R}$ of that data space, that is, $\pi_i(x) = x^i$. Similarly, $\pi_i \times \pi_j$ is the 2–D orthogonal projection $\mathbf{R}^n \rightarrow \mathbf{R}^2$ defined by $\pi_i \times \pi_j(x) = (x^i, x^j)$.

Definition 1 (2–D View) A view v is a 2–D orthogonal embedding of X to the (i, j) coordinates with $v = \pi_i \times \pi_j(X)$.

Second, we define the clustering algorithm and the resulting class structure in n–D. Clustering is the process of finding a partitioning of the data into homogeneous groups.

Definition 2 (Class Structure in n–D) Let X be a n–D data space $X \subseteq \mathbf{R}^n$. Let τ be an external source that labels data points as belonging to some classes. Then $C = \tau(X) = \{c_1, \cdots, c_m\}$ is the set of m classes called the class structure of X. Each class consists of the subset of the data space assigned to that class, and each data point is assigned to a class, thus

$$C = \tau(X) = \{c_1, \cdots, c_m\} \text{ and } \bigcap_i c_i = \emptyset \text{ and } \bigcup_i c_i = X$$

and $clabel : X \longrightarrow \mathbf{N}$ with $clabel(x)$ is the associated class label of each data point $x \in X$.

Note that in general $C(X)$ can be generated by any algorithm that classifies data, or by a-priori semantic information that divides the data points into categories. In our scenario a clustering algorithm or a supervised classification method is an external data preprocessing step that assigns labels to data points. By belonging to some class we mean to include both the simple situation of data points clustering around a cluster center, and the more general situation in which data points cluster around curved manifolds.

3.1. Basic Concept

We call a view v consistent with $C(X)$ when the m classes of $C(X)$ are mapped to regions in $v(X)$ that are visually separable. Note that in contrast to a (non)-linear cluster-preserving

transformation that seek to minimize or maximize various class separability criteria, class consistency is a computable measure that scores the utility of 2–D orthogonal projections to visually preserve a given class structure which can be used to choose the best views in a large matrix of scatterplots.

Definition 3 (Consistent View $v(X)$) Let X be a n–D data space $X \subseteq \mathbf{R}^n$. Let $C = \tau(X) = \{c_1, \cdots, c_m\}$ be the class structure of X with m classes.
We call a view $v(C)$ consistent with C iff

$$\forall x' \in v(X) \forall p' \in nbh(x') : clabel(p') = clabel(x') \quad (1)$$

with x' is the 2–D projection of a data point x, $nbh(x') = \{p' \in v(X) | d(x', p') < \varepsilon\}$, and d denotes a metric defined in X.

The level of consistency of a view depends on the definition of the threshold ε in the neighborhood function $nbh(p')$ which depends on the application scenario and task at hand. In the following section we propose methods to calculate the class consistency of a given view $v(X)$.

4. Class Consistency Algorithms

In this section we propose two methods for calculating class consistency, the centroid distance metric and distribution consistency. The distance metric can be used as a very efficient method to compute the class consistency of a view $v(X)$ if the class structure $C(X)$ describes convex classes. Distribution consistency is more general, but more expensive to compute. Since class consistency characterizes the extent to which the classes are preserved in a 2–D view, class consistency utilizes similar criteria used in traditional clustering algorithm to compute consistency scores. Note that in contrast to clustering algorithms, class consistency is meant to measure the utility of a given 2–D view to faithfully convey a given class structure to the user. A natural question is to utilize alternative clustering measures such as graph cuts in class consistency, which is left for future work.

4.1. Distance Consistency

Partitioning clustering algorithms such as k-means are widely used in data analysis. These methods create in general convex cluster models, because they aim to partition the data space into k clusters in a way that the quadratic distance of all cluster members to the centroid (the scatter around the centroid) is minimized. More precisely, we can observe that the distance between a cluster member and its centroid is minimal in comparison to all other centroids. We call that the centroid distance **CD**.

Definition 4 (Centroid Distance **CD**) Given a data space $X \subseteq \mathbf{R}^n$ and a class structure $C(X)$ defining m classes. Let c_i be a class and $centr(c_i)$ its centroid, and let x be $x \in X$ with $clabel(x) = i$. **CD** describes the property of class members that the distance $d(x, centr(c_i))$ to its class centroid should be always minimal in comparison to the distance to all other centroids, thus

$$d(x, centr(c_i)) < d(x, centr(c_j)) \; \forall j : 1 \leq j \leq m; j \neq i \quad (2)$$

and d denotes a metric defined in X. $\mathbf{CD}(x, centr(c_i)) = true$ denotes that the centroid property for x and its centroid $centr(c_i)$ is fulfilled.

The centroid distance **CD** is a good measure for the compactness and separation of classes in high-dimensional data spaces, and a low dimensional embedding capturing this basic property should also show separated classes. Note, if the embedding of two class centers in $v(X)$ just differs by a small ε, then the centroid distance property is always violated. We can use the centroid distance as an efficient measure for calculating consistency for given 2–D orthogonal projections. The idea of our consistency algorithm is to measure how well **CD** is preserved in a 2–D orthogonal projection. We evaluate a view by computing the percentage of data points for which **CD** is violated. *Distance consistency* will therefore be defined as the classification error of class members using **CD**.

Definition 5 (Distance Consistency **DSC**) Let $X \subseteq \mathbf{R}^n$ be a n–D data set with k data points. Let $C(X)$ be a class structure of X defining m classes $C(X) = \{c_1, \cdots, c_m\}$. Let c_i be a class and $centr(c_i)$ its centroid in $C(X)$. Let $clabel(x)$ be the class label of a point $x \in X$. Let $v(X)$ be a 2–D view of X, then distance consistency $\mathbf{DSC}(v(C))$ is defined as the classification error

$$\mathbf{DSC} = \frac{\left|x' \in v(X) : \mathbf{CD}(x', centr'(c_{clabel(x)})) \neq true\right|}{k} \quad (3)$$

with x' is the 2–D projection of the data point x and $centr'(c_i)$ is the 2–D projection of the centroid of class c_i.

We normalize the classification error to improve interpretability (score between 0 and 100). In practice, the number of clusters generated by a clustering algorithm is rather small relative to the number of data points. We only compute the distances of each point to this small set of cluster centers, and additionally we terminate the computation if we find a center that violates the centroid distance **CD**. Because of this property the computation time is roughly $O(k)$.

To demonstrate the usefulness of our measure we chose the pre-classified UCI [NHBM98] wine data set which has 3 distinct clusters defined by 3 different kinds of wine and 13 attributes describing their chemical properties. Figure 2 shows a well and poorly rated view of the wine data (the class structure is visualized using 3 different colors). In the well rated view (Figure 2 (a)) all 3 distinct clusters are separated. In contrast to the well-rated view in Figure 2 (a), the poorly rated view (Figure 2 (b)) completely merges the green class with the red and blue classes. In Figure 2 (b) it is not clear that there are 3 classes in n–D.

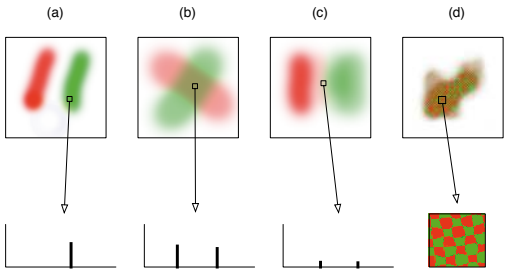


Figure 3: ***Basic Idea of Distribution Consistency – Top row****: hypothetical spatial distributions of projected data, with two classes represented as red and green.* ***Bottom row****: hypothetical histograms showing the proportion of data of each class in a small region such as a pixel. (a) The classes are clearly separated. (b) Classes are overlapping, and the histogram has higher entropy as a result. (c) Classes are overlapping in the indicated region, but the amount of data is small. The contribution of this region is weakly weighted. (d) Classes are spatially interleaved on a fine scale. Although each individual pixel contains only one class of data, the distribution has low distribution consistency when class proportions are estimated over a small window.*

4.2. Distribution Consistency

In this section we propose an extension of our distance consistency approach to accommodate more general spatial distributions that cannot be characterized as compact classes.

First consider a small region such as a single pixel (the size of the region will be reconsidered below). If the region contains data from only one class, consistency is completely satisfied. If the region contains an equal mixture of data from all classes, the region is least satisfactory according to this criterion.

Definition 6 (Entropy as a Measure of Randomness) Let $C(X) = \{c_1, \cdots, c_m\}$ be a class structure of a high-dimensional data space $X \subseteq \mathbf{R}^n$ describing m classes. Calling $p_c \equiv p_c(x,y)$ as the number of data points of class $c \in C(X)$ in the region centered at screen location x,y, the entropy of the class data probability density within the region

$$H(x,y) = -\sum_{c \in C(X)} \frac{p_c}{\sum p_c} \log_2 \left(\frac{p_c}{\sum p_c}\right) \tag{4}$$

is a measure of consistency violation, having minimum value zero if the region contains data from only one class (Figure 3 (a)), and maximum value $\log_2 m$ if all m classes are mixed equally (Figure 3 (b)).

This measure could be integrated over the whole image. However, doing so would weight regions equally, regardless of the amount of data they contain. Arguably, it is more important to be consistent in regions that contain more data. Thus, we weight the measure according to the amount of data in the region, $p(x,y) \equiv \sum_{c \in C(x)} p_c$ (see Figure 3 (c)).

Definition 7 (Distribution Consistency **DC**) Let $C(X) = \{c_1, \cdots, c_m\}$ be a class structure of a high-dimensional data space $X \subseteq \mathbf{R}^n$ describing m clusters. Let $v(X)$ be a 2–D view of X then distribution consistency $\mathbf{DC}(v(X))$ is a integrated and weighted measure with

$$\mathbf{DC} = 100 - \frac{1}{Z} \sum_{x,y} p(x,y) H(x,y) \tag{5}$$

The $1/Z$ is a normalizing constant chosen to improve interpretability. We choose $100/(\log_2(m) \sum_{x,y} \sum p_c)$ to give a score between 0 and 100.

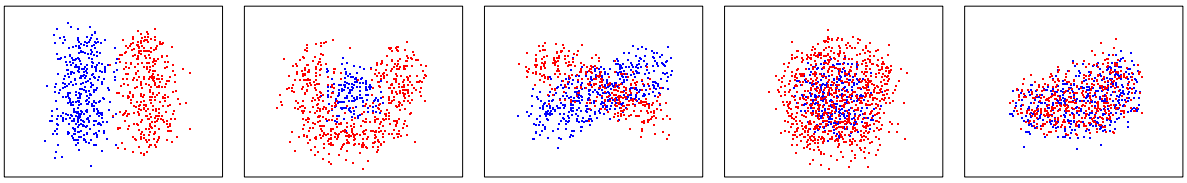
The performance of this measure on some difficult two-class synthetic distributions is shown in Figure 4. Note in particular Figure 4 (b), showing non-convex distributions that could not be handled with our earlier algorithm, and (d), showing a concentric (equal center, differing variance) distribution with partial but not complete overlap.

The region over which p_c is defined should be reconsidered now. If this region is a pixel, as suggested above, the measure will attempt to select views where individual pixels are consistent, but will allow pixels representing different classes to be intermixed arbitrarily. It is usually preferable however to have pixels of a single class grouped together, at least to the extent possible without violating other considerations. This will discourage "interleaved" data patterns such as in Figure 3 (d) (although such projections may be rare, note that grid-like data patterns do commonly arise in visualizations of network intrusion scans as different hosts and ports are accessed in sequence). The measure is altered to consider this by defining $p_c(x,y)$ to be the "amount" of data in a larger region σ centered at x,y, more specifically, the integral of the projected data under a weighting kernel of width σ. The choice of kernel width is an issue. In our case however, the kernel width has a direct interpretation as the size of a region over which classes should (preferably) not be mixed. The desired kernel width can thus be specified interactively with a slider.

Distribution consistency is relatively insensitive to the choice of σ except in the case of interleaved patterns such as Figure 3 (d). To demonstrate this, an interleaved pattern similar to Figure 3 (d) was added to the set in Figure 4. The following table shows the distribution consistency rankings for the patterns from Figure 4, with the relative ranking of the new pattern indicated numerically (from left to right):

$\sigma = .5\%$	100	(a)	(b)	(c)	(d)	(e)
$\sigma = 5\%$	(a)	(b)	(c)	(d)	24	(e)

It can be seen that with σ set to 5% of the image width, the interleaved pattern is rated as having low consistency (24), but the ranking of the other patterns is unchanged.



(a) 99 (b) 74 (c) 51 (d) 29 (e) 15

Figure 4: ***Rating of several synthetic patterns by distribution consistency*** *– the distribution consistency score is indicated below each figure. From left to right: (a) separate distributions, (b) separate non-convex distributions, (c) distributions partially overlap, (d) concentric distributions (same center, different variance), (e) identical distributions.*

5. Selecting Good Views in Large SPLOM

The challenge in exploratory data analysis is to find the highly revealing views of a high-dimensional data space. We demonstrate the benefit of class consistency for the interactive exploration of classes in large matrices of scatterplots. For this purpose, we integrated our class consistency measures into an exploration system called *Class Explorer*.

To demonstrate the usefulness of our class consistency measures we chose two pre-classified data sets. The UCI [NHBM98] wine data set which has 3 distinct clusters defined by 3 different kinds of wine and 13 attributes describing their chemical properties. The WHO HIV data set [Wor08] consists of 159 attributes describing socio-economic properties of 194 member countries such as birth attended by skilled health personal, life expectancy, access to safe water sources etc. The member countries are classified into 6 HIV risk groups.

Figure 5 shows typical exploration scenarios. In Figure 5 (a) it is difficult to manually detect good views to the wine data in reasonable time; even for moderate numbers of attributes. Our system computes the class consistency for each scatterplot, and the user can define class consistency thresholds via interactive sliders to fade out 'poor' views. After the consistency threshold is set to 80 many irrelevant views are faded out. It is easy to see that scatterplots which show all 3 clusters separated are detected as good views. The user may now analyze the remaining views, by selecting them for detailed analysis. Additionally, the user can interactively navigate through a ranking of views according to their class consistency, to analyze highly ranked scatterplots.

Figure 5 (b) shows a more reasonable scenario. The mapping of the 159 dimensions of the WHO data space into 2–D scatterplots results in over 12.000 unique views to the 6 HIV risk groups. An analyst typically inspects views till interesting patterns are found. Views that are cluttered or where the clusters mix provide little insight and are often considered uninteresting. Clearly, a human analyst cannot afford to look at every scatterplot in that huge SPLOM to explore mutual relationships of HIV risk groups because of his/her limited attention. Again, after the consistency threshold is set to 80, nearly 97% of the scatterplots are faded out. Figure 5 (b) shows a small part of the SPLOM of the 159-dimensional WHO data set. Scatterplots with low consistency scores are faded-out, and even the distribution of highlighted views across the SPLOM can reveal relations. In the WHO's SPLOM, many rows exclusively contain views with high consistency scores. A closer look at the dimension of one of these rows surprisingly shows that total expenditure on health as percentage of gross domestic product separates high-risk and low-risk cluster well. Besides this filtering step our method allows to rank views from high to low consistency values as shown in Figure 5 (b).

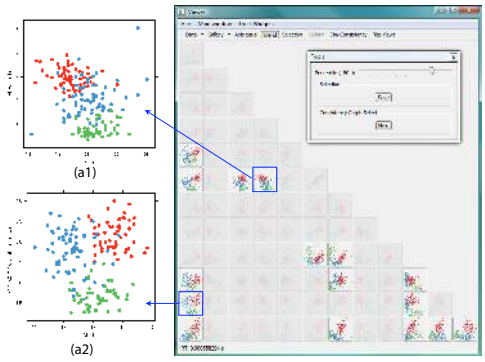
6. Evaluation

6.1. Consistency on Different Data Sets

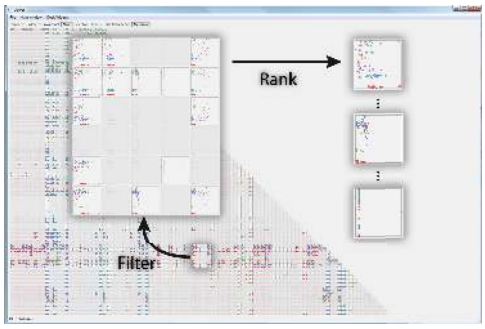
To evaluate our technique, we applied the consistency measures to a number of different data sets, including Iris, Wine, and Boston Housing data sets from the UCI repository [NHBM98], synthetic data sets, and unclassified data sets as well. For classified data, the consistency measure ranks how consistently the high dimensional classes are represented in the 2–D embeddings. For unclassified data we applied a clustering algorithm to generate a high dimensional class structure, and applied the consistency measure to analyze the consistency of the 2–D projections.

The max and mean distance consistency for all data sets are shown in Figure 6. The left figure shows that the number of consistent views decreases with increasing number of dimensions but our distance consistency measure still identifies a number of good views. For the Iris data for example, which is fairly simple since one of the three cluster can be linearly separated, our approach rated the views on average with **DSC** $= 90$.

For the Boston Housing data set we experimented with different numbers of classes. For this data set, the mean consistency decreases with increasing number of classes due to the decreasing separation between classes as shown in Figure 6. However, our measure still identifies good views with consistency with more than **DSC** $= 70$ as shown in the right figure. In general, these experiments show that our consistency measure is able to identify views that reveal the class structure in n–D.



(a) Wine (13 Dimensions)



(b) WHO (159 Dimensions)

Figure 5: ***Interactive Selection of good scatterplots*** *– interactive threshold sliders to fade out poor views supports to find good views interactively (a) . In this example, views below* **DSC** $= 80$ *are faded out. The projection of dimensions* $(1, 11)$ *for example has a high consistency of* **DSC** $= 86$*, as shown in (a2). (b) In the WHO example, views below* **DC** $= 80$ *are faded out. Many irrelevant views are faded-out and the number of views to look at can be interactively reduced to a manageable size.*

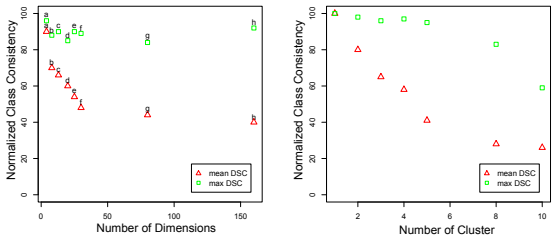


Figure 6: ***Normalized max and mean distance consistency for Iris (a), Olive (b), Wine (c), Boston Housing (d), Bulgaria Health [Bul08] (e), Health [NHBM98] (f), Artificial (g), WHO (h)*** *– it shows that the number of good views decreases but our technique identifies a number of good 2–D views for these data sets (left). For the Boston Housing data, the consistency decreases with increasing number of clusters due to the decreasing separation between clusters (right).*

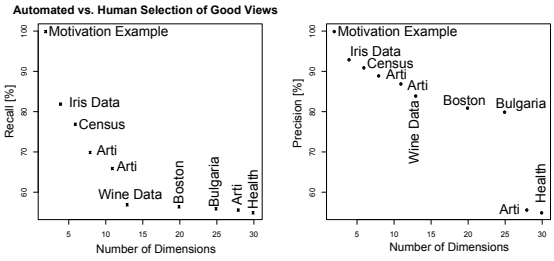


Figure 7: ***Precision and Recall*** *– even for a large number of dimensions the automatically detected good views correlates to over* 50% *with people's judgement of good views. Additionally we see that our distance consistency violates the human understanding of a good view only in a small number of views.*

6.2. Comparison with human judgement

We performed a small experiment to show the performance of our automated consistency measure in comparison to the human selection of good views. We asked 10 people from the graphics laboratory at Stanford to select 5 good views in different scatterplot matrices. We ran our experiments on a number of real-world and artificial data sets and the size of the matrices of scatterplots varied from small (4 and 8 dimensions) to very large (13 and 30 dimensions). We computed recall and precision to demonstrate the performance of our distance consistency.

Figure 7 shows the result of this experiment. The left figure shows that the performance of distance consistency is clearly related to the good views selected by humans. Furthermore, even for a large number of dimensions (hundreds of views) the automatically detected views are consistent with at least half of the sample population's judgments of what a good view is.

The right figure shows the effectiveness of our distance consistency. We can see that distance consistency finds some good views that are not selected by the user. We inspected these views and made the following two observations. First, human viewers have little preference when shown views differ in consistency by about 5% or less as rated by our measure, so the choice between fairly similar views is somewhat arbitrary. The second observation is that human observers may simply fail to notice every good view in datasets with more than a handful of plots. We can also see that even for a large number of dimensions our distance consistency detects almost all good views selected by the human, and therefore is in line with human judgement.

6.3. Comparison of consistency methods

Rather than repeating all these evaluations on the distribution consistency measure, we chose to simply examine the correlation between this measure and the distance consistency measure. The correlation between these measures is reasonably strong and is not sensitive to the kernel width (σ) parameter in the optimal region. This table shows the correlation as a function of σ for the *Wine* and *Iris* data:

σ	correlation (wine)	correlation (iris)
.03	71	81
.04	71	84
.05	70	86

In summary, our experiments show that our consistency measures are in line with human selection of good views.

7. Summary and Conclusion

In this paper we introduced class consistency as a criterion for automatically ranking and selecting good views to a class model from among the numerous possible projections of a high-dimensional data set. Class consistency characterizes the extent to which the class neighborhood structure in the high-dimensional data is preserved in a low-dimensional view. This method can be applied to data with preexisting categorical labels, or to data that has been organized into classes with a clustering algorithm.

Two computable measures of consistency were presented. The first, distance consistency, is easy to implement and is well suited for data with convex clusters. We found that this measure is correlated with people's preferred views of a variety of real world data sets. The second measure, distribution consistency, is more general and can assess non-convex and interleaved data distributions. We compared these two measures on a variety of data sets and found that they were highly correlated. The use of these consistency measures can reduce or eliminate the need for the analyst to manually search among a large number of data projections.

One issue that became apparent during our studies is that with increasing number of dimensions or clusters it is harder to find views that are highly consistent. The chances that clusters mix increases as the dimension and the number of clusters increase. This is a serious problem in real-world data analysis, in which case the analyst might consider other visualization techniques. Thus, class consistency can be used as a warning sign that suggests other techniques should be tried.

In this paper, we only considered scatterplot matrices. A natural question is whether these ideas can be applied to other types of visualizations. We could consider alternative projections or embeddings, or completely different visual metaphors. Consistency is a very powerful idea and could be generalized in many ways.

8. Acknowledgment

The authors thank the graphics lab people at Stanford for their great support in our user testing. Collaboration with Jörn Schneidewind led to useful insight. This work was supported by the Max Planck Center for Visual Computing and Communication.

References

[Asi85] ASIMOV D.: The grand tour: a tool for viewing multidimensional data. *SIAM Journal on Scientific and Statistical Computing 6*, 1 (1985), 128–143.

[Ber84] BERTIN J.: *Semiology of Graphics*. The University of Wisconsin Press, 1984.

[Bul08] BULGARIAN NATIONAL CENTER OF HEALTH INFORMATICS: Health Indicators of Bulgaria (http://212.122.183.76/dps/index.php), last accessed 03/2008.

[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.

[DMS98] DHILLON I. S., MODHA D. S., SPANGLER W. S.: Visualizing class structure of multidimensional data. In *Proceedings of the 30th Symposium on the Interface: Computing Science and Statistics* (1998), vol. 30, Interface Foundation of North America, pp. 488–493.

[Fri87] FRIEDMAN J. H.: Exploratory projection pursuit. *Journal of the American Statistical Association 82*, 397 (1987), 249–266.

[FT74] FRIEDMAN J. H., TUKEY J. W.: Projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers 23*, 9 (1974), 881–890.

[Har75] HARTIGAN J. A.: Printer graphics for clustering. *Journal of Statistical Computing and Simulation 4*, 3 (1975), 187–213.

[KC04] KOREN Y., CARMEL L.: Robust linear dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics 10*, 4 (2004), 459–470.

[KSA04] KEIM D. A., SIPS M., ANKERST M.: Visual data-mining techniques. In *Book Chapter in: Visualization Handbook*, Johnson C., Hansen C., (Eds.). Elsevier Science Publishing, 2004, pp. 813–825.

[NHBM98] NEWMAN D., HETTICH S., BLAKE C., MERZ C.: UCI repository of machine learning databases (http://www.ics.uci.edu/∼mlearn/mlrepository.html), 1998.

[SS05] SEO J., SHNEIDERMAN B.: A rank-by-feature framework for interactive exploration of multidimensional data. *Palgrave Macmillan Information Visualization 4*, 2 (2005), 96–113.

[TT85] TUKEY J. W., TUKEY P. A.: Computing graphics and exploratory data analysis: An introduction. In *Proceedings of the Sixth Annual Conference and Exposition* (1985), National Computer Graphics Association, pp. 773–785.

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

[Wor08] WORLD HEALTH ORGANIATION: WHOSIS WHO Statistical Information System (http://www.who.int/whosis/en/index.html), last accessed 03/2008.