DOI: 10.1111/j.1467-8659.2012.03126.x
Eurographics Conference on Visualization (EuroVis) 2012
S. Bruckner, S. Miksch, and H. Pfister
(Guest Editors)
Volume 31 (2012), Number 3

Semi-Supervised Dimensionality Reduction based on Partial Least Squares for Visual Analysis of High Dimensional Data

Jose Gustavo S. Paiva[1,3], William Robson Schwartz[2,4], Helio Pedrini[2] and Rosane Minghim[1]

[1]USP, Sao Carlos, Brazil, [2]UNICAMP, Campinas, Brazil, [3]UFU, Uberlandia, Brazil, [4]UFMG, Belo Horizonte, Brazil

Abstract

Dimensionality reduction is employed for visual data analysis as a way to obtaining reduced spaces for high dimensional data or to mapping data directly into 2D or 3D spaces. Although techniques have evolved to improve data segregation on reduced or visual spaces, they have limited capabilities for adjusting the results according to user's knowledge. In this paper, we propose a novel approach to handling both dimensionality reduction and visualization of high dimensional data, taking into account user's input. It employs Partial Least Squares (PLS), a statistical tool to perform retrieval of latent spaces focusing on the discriminability of the data. The method employs a training set for building a highly precise model that can then be applied to a much larger data set very effectively. The reduced data set can be exhibited using various existing visualization techniques. The training data is important to code user's knowledge into the loop. However, this work also devises a strategy for calculating PLS reduced spaces when no training data is available. The approach produces increasingly precise visual mappings as the user feeds back his or her knowledge and is capable of working with small and unbalanced training sets.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Display algorithms

1. Introduction

Dimension reduction is an important task involved in data analysis in general, and in particular for data sets that reach a high number of dimensions or attributes. It is applied in a visual analysis pipeline usually at the start of the process of visual mapping, whereby a method, such as Principal Component Analysis (PCA), is employed to find principal directions that recombine coordinates in new and fewer dimensions. To allow visualization of multidimensional data, various dimension reduction techniques have been employed to map from multidimensional to bidimensional spaces by reducing the dimension of the original space to two.

Two drawbacks of available projections are the inability to model the transformation in a way that it can be applied to data sets other than the ones used for the original data mapping, and also the difficulty in considering user's knowledge to influence the layout.

Partial Least Squares (PLS) [Wol85] is a highly effective technique that is usually employed for many different tasks involving automatic data segregation or regression, taking as input a labeled training set that is small in regards to the complete data, and can be unbalanced. For visual analysis purposes, these capabilities can be valuable.

In this paper, we propose a novel mapping process for visual analysis purposes that improves available strategies, by allowing users to input their knowledge into the dimension reduction process by means of the training set, and, from that, being able to create a reusable model for improved dimension reduction and visualization of larger data sets. Although labeling of the training set is an important step of the process, we also develop a strategy for PLS reduction in cases where no labeling of any part of the data set is available.

The following sections describe the background in 2D mapping techniques from multidimensional spaces, the basic concepts of PLS, our approach to visual mapping using PLS and, finally, the results with various text and image collections as well as a comparison with other approaches for dimension reduction purposes.

2. Dimension Reduction for Visualization and Visual Mining

In visual analysis applications, dimension reduction is sometimes achieved by selecting relevant attributes of the data set from which multi-attribute visualizations are generated. This is the case of the VHDR [YWRH03] and DOSFA [WPW03] approaches, that, based on measures between different data dimensions, support user's selecting and ordering dimensions.

The main motivation for selecting instead of combining variables for visualization is to maintain the meaning of the attributes themselves during exploration. However, in many applications such as visual analysis of text and image sets, the number of attributes easily reach hundreds or thousands, which makes attribute-based exploration unfeasible. Additionally, dimension reduction can improve results in mining as well as visual mining applications [TLKT09]. In such cases, it is important to find latent spaces with manageable number of dimensions that manage to represent well or to improve data spaces.

A second role played by dimension reduction is to provide a mapping to visual spaces by reducing original dimensionality directly to 2D or 3D spaces. When applied to that purpose, they are frequently called multidimensional projections, or simply projections. Depending on the goal of the mapping, projections either aim at achieving discriminability of groups or reproducing dissimilarity relationships existing in original spaces.

Projection techniques are usually based on Multidimensional Scaling (MDS) methods, with various mathematical foundations, such as spectral decomposition from transformations on similarity matrices [Tor65,BN03,KCH02]. Some spectral-based methods, such as ISOMAP [TdSL00], can also deal well with non-Euclidean distances, which is useful for visualizing data generated from various feature description methods. Although representation properties and processing time have been largely improved in later formulations of MDS algorithms, the global aspect of these methods impair their use in applications where it is important to capture or emphasize near neighborhoods.

Some spectral-based methods, such as Locally Linear Embedding (LLE) [RS00], Landmark MDS (LMDS) [dST04] and Pivot MDS [BP07], can actually capture local aspects of data. However, the global eigendecomposition scheme in these methods prevents them from being used in applications that redefine local relationships since an update forces recalculations that are global to the model they produce.

Some techniques have been designed to deal with samples of data to cope with high computational costs of decompositions. One example is Sammon projection [PdRDK99], that is meant to improve dimension reduction and mapping based on optimization methods [Kru64,BBKY06], typically heavy in computational cost. Although this idea could be used to embed user's knowledge by first mapping a sample set and then adjusting the others, its formulation does not produce a final model to be reused in other data set mappings.

There exist local projections designed for visualization purposes. Paulovich et al.'s [PNML08] Least Squares Projection (LSP) employs a force-based scheme to first position a subset of the samples, mapping the remaining instances through a Laplace-like operator. It results in a large linear system that is strong in local feature definition. Its derived methods, such as Piecewise Laplacian-based Projection (PLP) [PEP*11] and Local Affine Multidimensional Projection (LAMP) [JCC*11] have progressively managed to allow redefinition of the mapping matrix under user's intervention over a first mapping of sampled instances. These have more local formulations and are more flexible as far as rearranging the visual mapping. However, they are not capable of using a mapping to project data sets other than the one used to produce the mapping. Nor do they support user's choice of number of dimensions effectively.

The most common dimension reduction technique employed for data analysis in basically any field of science is by far the Principal Component Analysis (PCA) [Jol02], which finds principal directions in a covariance matrix calculated over the data dimensions. Regardless its high computational cost as a dimension reduction technique, PCA is very effective. It has also been used numerous times for mapping data visually into 2D or 3D. In such case, however, PCA performs worse, since choosing only the two first principal directions frequently fails to deliver proper data segregation.

Also based on the principle of finding latent spaces in feature data sets, Partial Least Squares (PLS) [Eld04, Wol85] has been recovered from the field of statistics to find many current uses in analysis of high-dimensional data, such as discrimination, feature selection, treatment of missing data problem and regression [BS07]. Some of the work on the large number of applications for PLS also employ MDS techniques (such as [LN06]) but they do not support visual analysis of one of the other, or allow for iterations or progressive refinement of the process. In this work, we show PLS to be a very flexible and precise tool for visual analysis of data sets by supporting user's feedback into the dimension reduction process, being capable of working with low number of samples, and supporting the reuse of the model to visualize increasing data sets. Reduced data sets improve original data sets regarding data discrimination.

Reduced data spaces can be visualized by MDS strategies as well as multiple axis visualization, such as Radviz [HGP99]. The reduced data set, can also be effectively visualized employing similarity trees, such as Neighbor-joining (NJ) trees, which have been previously used as support for visual classification tasks [PFCP*11].

3. Partial Least Squares for Visualization of Multidimensional Data Sets

Partial Least Squares is a class of statistical methods used to model relations between sets of observed variables by the estimation of a low dimensional latent space. Its goal is to estimate a low dimensional space that maximizes the separation between samples with different characteristics, causing samples from the same class to be clustered in the latent space. Here we describe its uses, its formulation and how to apply PLS for data classification and dimension reduction.

The underlying assumption of PLS methods is that the observed data is generated by a system or process which is driven by a small number of latent variables. This way, it reduces the number of dimensions prior to the estimation of the regression coefficients, so that the influence of high dimensional noisy samples is reduced, thus improving discrimination results [Gar94].

PLS was created by Herman Wold in the 1970s [Gel88] and has been exploited in several areas, such as Chemometrics [BGJ*97, LGB*95], Bioinformatics [NR02, BS07] and Neurosciences [NOM*02]. Recently, PLS has been successfully applied to Computer Vision problems considering dimension reduction, regression and data classification [SKHD09, KHD11].

3.1. Dimension Reduction and Regression Based on Partial Least Squares

With the advantages of being designed to work in problems containing high dimensional data and very few samples [Gel88], PLS estimates latent variables as a linear combination of the original variables in a matrix X, composed of variables used to describe samples, and a matrix Y containing a set of response variables (when a single response variable is considered, a vector y is used instead). A description of the PLS decomposition and latent space estimation is given as follows.

For a problem with n samples described by d variables each, stored in a mean-centered matrix $X_{n\times d}$, associated to k response variables, stored in a mean-centered matrix $Y_{n\times k}$, PLS estimates a p-dimensional space ($p \ll d$) by performing the decomposition of X and Y into

$$X = TP^T + E \qquad \text{and} \qquad Y = UQ^T + F$$

where $T_{n\times p}$ and $U_{n\times p}$ are matrices containing the latent variables, matrices $P_{d\times p}$ and $Q_{k\times p}$ represent the loadings, and matrices $E_{n\times d}$ and $F_{n\times k}$ are the residuals. An approach to performing the decomposition above employs the nonlinear iterative partial least squares (NIPALS) algorithm [Wol85]. NIPALS estimates a set of projection vectors w_i ($i = 1,2,\ldots,p$), which are stored in a matrix $W = (w_1, w_2, \ldots, w_p)$, such that

$$[cov(t_i,u_i)]^2 = \max_{|w_i|=|u_i|=1} [cov(Xw_i, Yu_i)]^2 \qquad (1)$$

where $|w_i|$ and $|u_i|$ denote the 2-norm of vectors w_i and u_i, respectively. t_i and u_i represent the i-th columns of matrices T and U, and $cov(t_i,u_i)$ is the sample covariance between latent vectors t_i and u_i.

The NIPALS algorithm extracts the latent variables t_i and u_i iteratively. After each iteration, matrices X and Y are deflated by subtracting their rank-one approximations as

$$X_{i+1} = X_i - t_i p_i^T \qquad \text{and} \qquad Y_{i+1} = Y_i - t_i q_i^T$$

where X_i and Y_i are the data representation for the i-th iteration, where $X_1 = X$ and $Y_1 = Y$, and p_i and q_i denote the i-th columns of the matrices P and Q, respectively. After the extraction of p projection vectors, the p-dimensional representation of $X_{n\times d}$ is given by $T_{n\times p}$, which is used to extract the regression coefficients $\beta_{d\times k}$ as $\beta = W(P^TW)^{-1}T^TY$. Finally, the regression responses, Y_v, for a feature vector $v_{d\times 1}$ is obtained by $Y_v = \overline{Y} + \beta^T vS$, where $\overline{Y}_{1\times k}$ is the sample mean of each variable of Y and $S_{1\times k}$ is the standard deviation of the variables in Y.

As pointed out earlier, PLS performs dimension reduction in a supervised manner, differently from PCA. Due to its supervised nature, PLS estimates latent spaces that focus on the discrimination of the data. Therefore, the reduced space presents a better separation, which aids content-based visualization.

3.2. Data Classification Based on Partial Least Squares

According to the PLS formulation above, after the selection of the number of latent variables, referred to as *factors*, the NIPALS algorithm is applied to estimate a low dimensional representation of the original data.

3.2.1. One-against-all Classification

Aiming at maximizing the discrimination between C different classes, the one-against-all classification scheme estimates C PLS models considering single response variables [SGCD12]. This way, the response variable Y, represented by a matrix in Section 3.1, becomes a vector, y, and its entries have class indicators. In this work, we set +1 for positive samples (samples belonging to the class being modeled) and -1 for negative samples (remaining training samples). When a test sample is presented, it is projected to each model, resulting in a set with C responses (one per class) and the best matching class is associated to the model presenting the highest regression response.

3.2.2. Multi-class Classification

Differently from the one-against-all scheme, the multi-class creates a single PLS model containing multiple response variables. For a problem with C classes, the response variable Y, represented by a matrix in Section 3.1, has C columns, each one corresponds to one class and indicators variables are used to identify which samples belong to a

given class. In this work, the value +1 is set to $Y_{i,j}$ if the i-th sample belongs to the j-th class, otherwise, it receives 0. For a test sample projected onto the model, C responses are obtained and the best matching class is the one presenting the highest regression response.

Although the multi-class approach presents a faster latent space estimation compared to the one-against-all classification scheme, in general, the latter presents higher classification rates.

3.3. A Visual Analysis Approach Based on PLS Reduction

In this work, PLS is used in two ways for mapping a data set to low dimensional spaces targeting visualization. One way is to employ the conventional PLS dimension reduction, which considers the multi-class dimension reduction to p factors, using the matrix T, that is, the low dimensional representation of the data stored in matrix X (see Section 3.1), to generate a new reduced space with dimension p. The other way that its formulation may result in a low dimensional space is by using the responses in each class, estimated by the one-against-all classification, as coordinates for a projected sample in the reduced space. In this case, the reduced space has C dimensions.

In both cases, in order to visualize patterns in the original space through the reduced space, one can apply any of the available projection or point placement techniques mentioned in Section 2. To understand the distribution of the data over the final reduced dimensions, analysts can make use of any of the multi-axis visualization techniques available. We find that Radviz [HGP99] is useful in this case to help identify the contribution of particular latent dimensions in the data distribution, but tested with various others (see Section 4).

PLS needs as input a training set. However, we have also devised an approach to PLS mapping of unlabeled data sets. The approach to labeled and unlabeled data sets work as follows:

Approach 1 Data sets with labeled training sets.

1. Model creation phase: a PLS model is built from the labeled training set provided by the user. The user has the choice of using a multi-class or a one-against-all classification process.
2. Model application phase: the test data set is applied to the model created in the previous phase. The result is a reduced space that includes all points applied to the model, with dimensions either p or c depending on the choice of the classification process.
3. Visual mapping phase: the reduced model is projected onto the visualization plane using a point placement strategy.

Approach 2 Data sets with unlabeled training sets.

1. Label creation phase: the data is clustered attributing to each point the label of the cluster that belongs. The result is a data set labeled by clustering.
2. Model creation phase: it is the same as the model creation phase of approach 1, only the classes are now the cluster labels and the training set is sampled from the whole data set according to some sampling strategy.
3. Model application phase: it is also the same as in approach 1, only now the data set has no labels. The labels are assigned according to the winning class of the classification method.
4. Visual mapping phase: the reduced model is projected onto the visualization plane using a point placement strategy.

In either approach, a training set can be built by sampling a previously labeled data set. In our system and experiments, there are two forms of sampling. In the first form, the labeled data is clustered and an equal number of samples for each cluster is selected. Half of these are the closest points to each cluster centroid and the other half are the points in the cluster that are the farthest apart from the centroid.

The second form of sampling builds in the process a semi-supervised approach. From a distance-based initial visualization of the whole data set or a subset of it, the user manually chooses points from the data sets. In our system, the user can select those points in various different ways on the visualization itself. For the similarity tree visualization, for instance, points can be selected via a branch selection or by the area of a polygon.

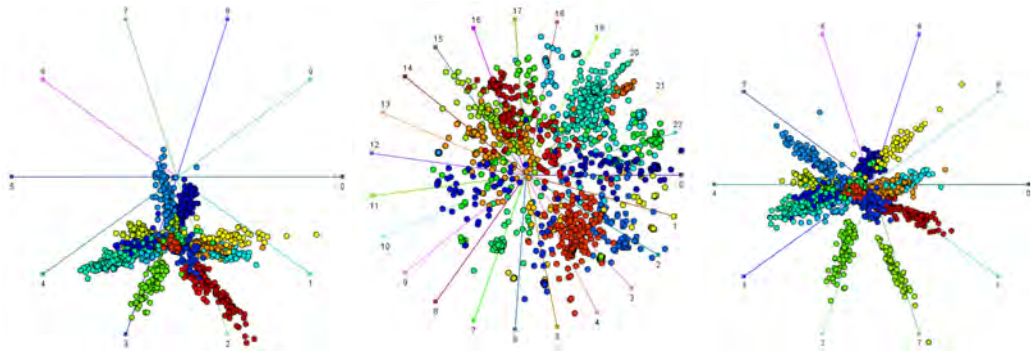
In fact, the clustering process necessary for unlabeled PLS mapping in approach 2 can also be done manually in our system by 'coloring' the whole data set group by group as a way to create an initial labeling.

The whole process can iterate until a proper sample and a proper model be built. The model can then be used to map any number of similarly described data sets. In our experience so far, not many iterations are necessary since the PLS methods produce very reliable models. The next section shows our analysis of the PLS mapping approaches.

4. Results

From the description of our methodology in the previous section, it can be seen that there are two parameters for PLS dimension reduction, i.e., a user-defined number of factors and number of classes determined by the number of classes (for labeled training sets) or number of clusters (for unlabeled training sets). From these parameters and the training set, the multi-class method will reduce the data dimension to the number of factors and the one-against-all (o-a-a) method will reduce the dimension to the number of classes.

For instance, Figure 1 shows a possible visual output for the data set NEWS. Figure 1(a) shows a visualization of its



(a) PLS map after multi-class reduction (b) PLS o-a-a reduction (c) Reordering of axes in (a)

Figure 1: *Examples of PLS dimension reduction of the NEWS 23 labeled data set with Radviz placement on latent dimensions. (a) is the multi-class reduction on 10 factors; (b) is o-a-a reduction on 23 classes, also 10 factors; (c) is another view of the space in (a).*

reduction by the multi-class strategy, shown by Radviz with the anchors representing the factors. Figure 1(b) shows the placement on the number of labels given by the o-a-a strategy. For a data set with 23 labels, 23 axes will be produced. Both strategies perform a considerable dimension reduction from the original 3731 dimensions of this data set. This view shows which dimensions are more involved in determining segregation of at least some of the classes present in the data. As it happens with Radviz, changing the order of axes will change placement of points. Since it is very fast, the user can experiment with different positions to understand the latent dimensions. Figure 1(c) shows the same reduction as in Figure 1(a), with order change in the axes.

In our tests, the choice of number of factors was always 10 for textual data sets. For image data sets, the number of factors was 8 for o-a-a and 5 for multi-class methods. We employed a k-fold cross-validation using the training samples to choose the number of latent variables, considering $k = 5$.

4.1. Data Sets and Test Setup

Table 1 presents details of the data sets employed in the evaluation tests.

Table 1: *Information on test data sets.*

Data set	Content	Classes	Items	Attributes
NEWS	RSS Feeds	22	1771	3731
ALL	Scientific Papers	8	2814	12201
COREL	Photographs	10	1000	150
ETHZ	Photographs	28	2019	3963

The NEWS data set was formed from 1771 RSS news feeds from BBC, CNN, Reuters and Associated Press, collected from their site between June and July 2011. From the text set, a feature space was created by removing stopwords and employing stemming. The coordinate of any particular point was determined by the *term-frequency-inverse-document-frequency* count. The result is a data set with 3731 dimensions. The 23 labels of the data set were assigned manually based on the perceived main topic of the news feed. The labels are unbalanced in number of points and there is high similarity of content between points labeled differently.

The data set named ALL contains abstracts of scientific papers in 8 areas of knowledge, with considerable part of common content across labels. This data set was collected from various sources and preprocessed similarly to the NEWS data set.

The COREL image collection[†] is composed of 1000 photographs that represent 10 specific subjects. Each image is represented by a vector of 150 SIFT descriptors [LW03]. The ETHZ image collection represents a subset of the ETHZ dataset [ELS*08], which provides photographs of different people captured in uncontrolled conditions, with a range of appearances. This collection is composed of 2019 images, divided into 28 labels forming unbalanced groups. Each image is represented by a vector of 3963 visual descriptors, combining Gabor filters, Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP) and mean intensity.

In order to verify the precision and feasibility of PLS mappings, we devised case studies that cover three situations: use of training sets selected by the user, use of training sets produced by clustering and sampling a labeled data set, and applying approach 2 (defined in Section 3.3) for unlabeled data. We also compared our approach to other dimension reduction techniques and tested various visual mappings for the reduced spaces.

[†] UCI KDD Archive, http://kdd.ics.uci.edu

Besides the visual output, the *silhouette coefficient* [TSK05] was used to evaluate the produced results numerically, since the main target of our approach is discriminability. The silhouette coefficient is a measure of cohesion and separation between groups of instances. Given an instance p_i, its cohesion a_i is the average distance between p_i and all other instances belonging to the same group as p_i. Its separation b_i is the minimum distance between p_i and all the other distances belonging to the other groups. The silhouette of a particular space or projection is given by the average of silhouette coefficients of all its n instances. Its formulation is given in Equation 2. Although the silhouette may not be the most appropriate measure to reflect grouping in projections from reduced spaces due to its inadequacy measuring clusters that are not round in shape, it is used here for assessment of group separation in the reduced spaces and also on visualization planes after 2D mappings.

$$S = \frac{1}{n}\sum_{i=1}^{n}\frac{(b_i - a_i)}{max(a_i, b_i)} \quad (2)$$

The silhouette coefficient varies between -1 and 1, with 1 meaning that groups are perfectly separated from one another. The distance measures employed were cosine for original textual spaces, Euclidean for original image spaces and Euclidean for all reduced spaces. We report detailed results for NEWS and ETHZ and summary results for data sets ALL and COREL.

4.2. Textual and Image Mappings

Textual data sets tend to produce sparse feature spaces whereas image collections produce denser data spaces. We tested PLS under both situations. In the next section, training sets are generated and then applied to unlabeled data. In order to verify precision, the labels are then painted on the visualizations. Then, in Section 4.2.2, the results for unlabeled data sets are given.

4.2.1. User Input and Sampling

The generation of training sets was done in two different ways, by allowing the user to select the samples or by automatically sampling a labeled data set (see Section 3.3). In our captions and tables, training sets manually defined by the user are identified by the suffix 'user' and sampled training sets by k-means clustering are identified by the suffix 'k-means', with k replaced by the proper number of clusters.

Table 2 shows numerical results of applying PLS reduction to the NEWS data set. The first line shows the silhouette coefficient of the original data set. The two following lines display the silhouette of the visualizations by LSP and NJ-tree as references. The subsequent lines present the silhouettes of reduced spaces after model creation with the training sets. We also show the silhouette of their projection using Radviz.

It can be seen from Table 2 that separability grows as the sample size increases. From the smallest sample set, PLS reduced spaces are improved twofold from original spaces in terms of silhouette coefficient.

Table 2: *Labeled reduction for NEWS data set.*

Sampling Size	Sampling Method	Silhouette reduced	Projection Technique	Silhouette
–	–	–	original	0.1374
–	–	–	LSP	0.0934
–	–	–	NJ-tree	0.0949
611	user	0.4052	Radviz	0.0612
863	user	0.6815	Radviz	0.1755
1169	user	0.7780	Radviz	0.4354
600	23-means	0.6187	Radviz	0.1035
800	23-means	0.5132	Radviz	0.1909
800	23-means multi	0.2799	Radviz	0.0537

Table 3 shows the same analysis using the ETHZ collection. Here, one can also notice a better separability of the reduced spaces that is proportional to the sample size growth and even greater than that reported for the NEWS data set. In that table, we also notice that Radviz had more difficulty in reflecting on the layout the silhouette of the data set, which is probably due to a larger number of points and classes.

Table 3: *Labeled reduction for ETHZ collection.*

Sampling Size	Sampling Method	Silhouette reduced	Projection Technique	Silhouette
–	–	–	original	0.0912
–	–	–	LSP	-0.0390
–	–	–	NJ-tree	0.1023
200	user	0.3622	Radviz	-0.1317
600	user	0.5457	Radviz	-0.0433
1000	user	0.6277	Radviz	-0.0555
200	28-means	0.4024	Radviz	-0.0777
600	28-means	0.5326	Radviz	-0.0648
1000	28-means	0.5915	Radviz	-0.0525

Figure 2 shows the precision of mappings from 863 samples with user defined training set by employing various 2D mapping strategies. Neighborhood Hit averages the number of neighbors for each point that belongs to the same class as that point. It can be seen that the NJ-tree is the best one to reflect class neighborhoods from reduced data. ISOMAP and LSP also perform well. From the plots, it can be seen that Radviz, in terms of class segregation, did not perform as well. That is due to its trend to place together points with same balance in coordinates, rather than similar coordinates.

Figure 3 shows three projections of the reduced NEWS data set produced using 2D multidimensional mappings and also one of the original data set. Using ISOMAP, groups are more dense, while Radviz tends to spread the groups more, with the parameter circle adjusted to the maximum coordinates of the data set. The others show good discriminability. The NJ-tree (Figure 3(c)) manages to reflect PLS separation of almost every class in the reduced space as well as levels of similarity within each class. Contrasting that with the tree built directly from the original attributes (Figure 3(d)), it can be seen that class neighborhoods of many points were resolved by the new reduced space.

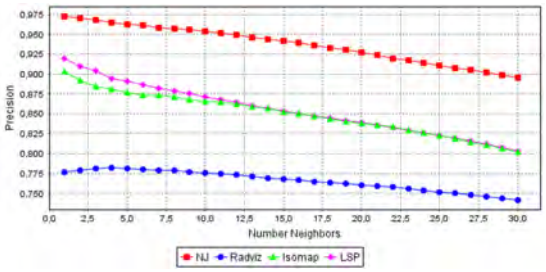
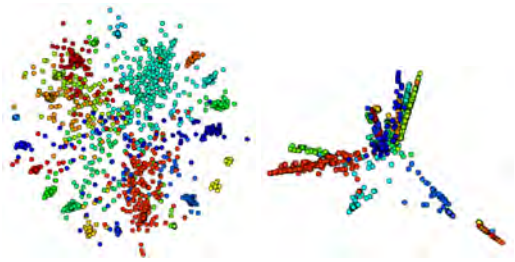
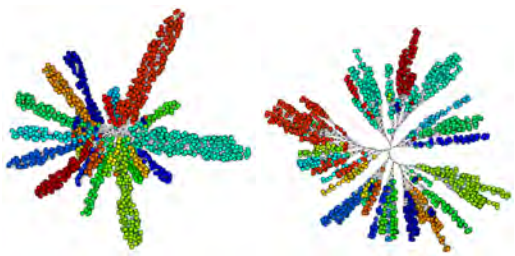


Figure 2: *Neighborhood Hit for projections of the NEWS data set, employing a user selected training set with 863 samples.*



(a) Radviz (b) Isomap



(c) NJ-tree (d) NJ-tree 3731 dim.

Figure 3: *(a) to (c) visualizations of the NEWS data set reduced by PLS to 23 dimensions; (d) NJ-tree of the data set with original dimensions.*

4.2.2. Unlabeled Datasets

Due to the nature of the PLS, training is necessary; therefore, labeling sets is a necessary task for the training set. To find out what could be done in case of such a training set is not available, we clustered the NEWS and ETHZ data sets and used the cluster label to produce a model using PLS. The reasoning behind it is that it is not necessary to know precise labels, but rather get proper samples that distinguish classes of points in order to obtain a PLS model with good segregation.

To cluster the unlabeled data set, we used two approaches. The first one clustered the data automatically by bisecting *k*-means (bkmeans) with 23 or 40 classes for the NEWS data set and with 20 or 28 classes for the ETHZ collection. In each cluster, samples were chosen as the closest and farthest from the centroid of the cluster. The second cluster approach required a manual classification procedure, supported by our system, that is done by projecting the data using the NJ-tree, which produces the same tree as Figures 3(c) and 3(d), except for the colors. From that, the user can label groups of points in the same branches by clicking on the top node of the branch or by drawing a polygon around a group that he or she intends to assign a label. We did that resulting in 24 classes (slightly different from the original 23) for NEWS data set and 12 classes (different from the original 28) for ETHZ collection.

Table 4 shows the result of applying this method for dimension reduction based on PLS using originally unlabeled training sets. The table shows two silhouettes for each reduction. One of them (Silhouette Reduced) is for the reduced space using the cluster label, and the 'Cross Silhouette' column is the silhouette of the reduced space considering the original label for every point, instead of the cluster label. It can be seen that the results are very satisfactory, with silhouettes varying from 0.07 to 0.41 for a data set with original silhouette 0.1374, in the case of NEWS data set. In addition, the one-against-all approach is more precise than multiclass.

Table 4: *Results of dimensionality reduction using unlabeled training sets.*

Data	Samples	Sampling Method	Reduction Method	Silhouette Reduced	Cross Silhouette
NEWS	original	–	–		0.1374
NEWS	800	bkmeans-23	multi	0.1718	0.1669
NEWS	800	bkmeans-23	o-a-a	0.4260	0.2388
NEWS	800	bkmeans-40	multi	0.1440	0.0705
NEWS	800	bkmeans-40	o-a-a	0.4173	0.1549
NEWS	800	nj-24	o-a-a	0.2913	0.3100
NEWS	all	bkmeans-23	o-a-a	0.9258	0.3529
ETHZ	original	–	–	–	0.0912
ETHZ	1000	bkmeans-20	multi	0.1060	0.1294
ETHZ	1000	bkmeans-20	o-a-a	0.3401	0.0191
ETHZ	1000	bkmeans-28	multi	0.1188	0.1014
ETHZ	1000	bkmeans-28	o-a-a	0.3172	0.0648
ETHZ	1000	nj-12	o-a-a	0.5236	-0.0884

Figure 4 shows the Neighborhood Hit of various projections from the unlabeled version of NEWS data set. Values are quite high, with the tree also performing best, followed by LSP.

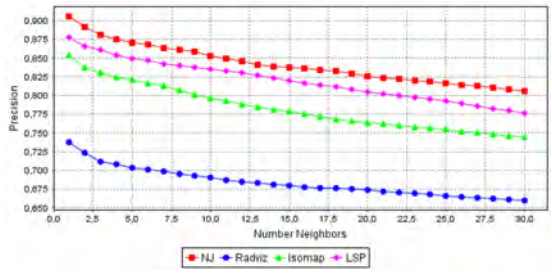


Figure 4: *Neighborhood Hit for projections of the unlabeled reduced NEWS data set, clustered with all data points.*

The data set named ALL was processed in the same way as the news data set. It is a very difficult data set to separate, since articles with very similar content have different labels and there are two classes that significantly dominate the data set while others are very small in numbers. The silhouette coefficient of 0.0350 for the original data gives a good idea of this type of distribution. For that data set, the silhouette coefficient of the reduction in the one-against-all method for 800 samples (less than a fourth of the data set) was 0.4926. For 1100 samples, it reached 0.5588. The NJ-tree for the reduced space has a silhouette of 0.2327, whereas for the original data it is 0.06.

The COREL image collection has originally a good separability of instances in their classes. The original silhouette for this data set is 0.1595. The dimensionality reduction using one-against-all method produces a reduced data set with silhouette coefficient of 0.5102, using 200 samples, improving considerably the original space. Using 500 samples, the silhouette coefficient is 0.5712. The values of silhouette for the ISOMAP projection are 0.1034 and 0.3780 respectively, before and after the reduction procedure. For NJ-trees, these values are 0.12 and 0.4002, also confirming the results obtained with ETHZ collection. Radviz cannot be applied on the original spaces due to space limitations for a large number of axes. For this example, the Radviz projection shows a silhouette value of 0.2759.

4.3. Times and Comparison with other Supervised Reduction Techniques

Tests with text files were run on a computer with Intel i5 processor with 2.3GHz and 6GB memory. For image examples, we have employed a notebook with Intel Core2 Duo processor, with 2.53GHz and 4GB RAM. We have implemented PCA plus three other dimension reduction techniques with supervision capabilities. Self-Organizing Maps would also satisfy the supervision requirement, but they proved to be extremely slow for reducing the large number of dimensions to a reduced space with more than three dimensions.

Table 5 shows computational times and silhouettes of reduced spaces obtained by PLS, as well as PCA, PivotMDS, ISOMAP and LLE. It can be seen that PLS performs similarly or better than PCA, depending on the strategy, and better than all the others. It is competitive with PCA in terms of time. The fastest ones do not perform as well.

Figure 5 shows the precision by Neighborhood Hit of six reduced spaces and for the original NEWS data set. The PLS reduced spaces employ the model created by sampling 800 labeled points shown in the previous section and a new one, with 510 samples chosen slightly more carefully, reduced both by multi-class and o-a-a. These user-defined 510 samples were very unbalanced within the 23 classes, with samples for each label varying from 7 to 57. The complete NEWS data set is also unbalanced, but in a different proportion to the samples. Results confirm the discriminability

Table 5: *Model creation times and silhouettes, compared to other supervised dimension reduction methods. PLS models are o-a-a, the slowest.*

Data Set	Reduction	Time	Silhouette		
			Reduced	LSP	NJ-Tree
NEWS	PCA-23	38 min	0.3163	0.0269	0.2189
NEWS	PLS 23-unlabeled	22 min	0.380	0.1210	0.27
NEWS	PLS-23-user-800	3 min	0.6815	0.1244	0.3456
NEWS	PLS-10-means-800	7 min	0.279	0.1363	0.2183
NEWS	PLS user-510	7 min	0.60	0.0665	0.3530
NEWS	LLE	2.5 min	-0,0195	-0,0127	-0,2491
NEWS	ISOMAP	11 sec	-0,2720	-0,3040	-0,2266
NEWS	PivotMDS	14 sec	-0,2062	-0,3340	0,1665
ETHZ	PCA-28	46 min	0.1039	-0.0674	0.0972
ETHZ	PLS-28-user-1000	12 min	0.6277	0.7928	0.5748
ETHZ	PLS-28-means-1000	18 min	0.6132	0.5107	0.5576
ETHZ	LLE	8 min	0.08131	-0.2085	0.0884
ETHZ	ISOMAP	32 sec	-0.0652	-0.2442	0.0
ETHZ	PivotMDS	48 sec	-0.1979	-0.3429	-0.1339

of PLS. PCA performs as well as PLS with the user 510 o-a-a set and sampling 800 set. All the other tested reduction techniques perform worse.

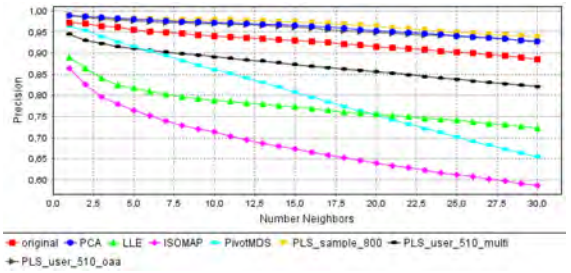
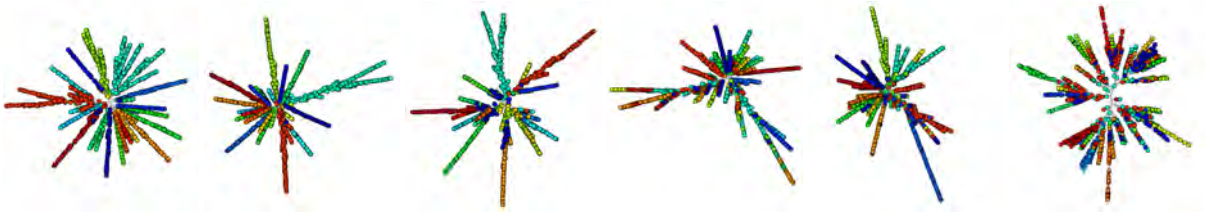


Figure 5: *Neighborhood Hit for the original and reduced NEWS data spaces. There is no statistical significance for the difference in precision between the spaces reduced by PCA, PLS unlabeled, and PLS by user training with 510 samples. They are statistically better than the original neighborhood hit, which is, in its turn, better than PLS multi-class from user samples, followed by PivotMDS, LLE and, finally, ISOMAP.*

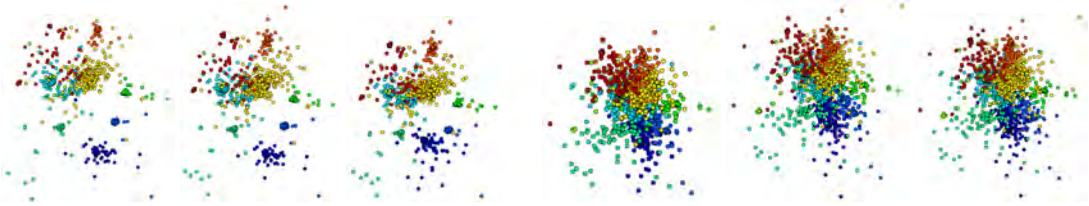
PLS times shown in the tables were spent mostly in the model creation phase of the PLS dimension reduction. After the model is generated, it is used to map any size of data set bearing the same features. This mapping from a prebuilt model is very fast. Loading and applying the model for ETHZ data sets took an average of 19s for one-against-all and 1.3s for multi-class. For the NEWS data set, it took 9.3s for one-against-all and 1.3s for multi-class. For the ALL data set, it took 4.5s for one-against-all and 0.6s for multi-class. The projections took different times with Radviz being very fast (2s at most) and the tree the slowest, but still taking few seconds.

In terms of visual quality of visualization, the silhouette is also a good measure of the visual improvement provided by reduced spaces. Figure 6 shows the NJ-tree layouts of the NEWS data set created from various of the reduced spaces mentioned above. An association between silhouette



(a) PCA (b) PLSoaa (c) PLSmulti (d) PivotMDS (e) LLE (f) ISOMAP

Figure 6: *NJ trees of the NEWS data set generated from reduced spaces.*



(a) 796 points (b) 1237 points (c) 1520 points (d) 1968 points (e) 2402 points (f) all 2814 points

Figure 7: *Progressive application of the ALL data to the model created by sampling 1200 points using the clustering strategy.*

and quality of visualization can be observed, related with grouping, on the same branches of individuals with the same labels.

Finally, one of the important motivations of the method is precisely that it is able to use a model to map growing data sets. Figure 7 shows a series of mappings from a previously calculated model for different sizes of the ALL data set, using one-against-all strategy. It shows that, as the data set increases in size, the mappings of certain groups of points are kept in the same region. This allows the user to maintain the mental model of the PLS projection. A model can be used until it is no longer adequate to represent the changes in the data set, in which case the user can adjust the model again to comply with possible changes in class distributions.

A Java API for the PLS techniques as well as a system that implements the visual approach and all the data sets used in this section are available at `http://infoserver.lcad.icmc.usp.br`, following the link to the Tools section.

5. Conclusions and Future Work

Although PLS is not a new technique for discrimination analysis, its use in the context of visual analysis is novel and, as it has been shown in this work, PLS is a very powerful tool that supports various important tasks for visual data mining, such as dimension reduction, classification and manipulation of growing data sets.

The detailed analysis presented here shows the high precision of PLS dimension reduction both to create a model from a labeled data set to be applied effectively for larger data and to generate effective models for unlabeled data sets. The approach based on user training sets aids the user to build and change his or her view of the data adapting the system to adjust according to acquired model. The extensive analysis presented is meant to offer evidences of the flexibility of PLS in various visual analysis contexts.

Within the visual mapping framework, we have shown that reduced spaces can be successfully visualized by using multidimensional projections or similarity trees, reflecting proper improvement of the data space provided by the dimension reduction.

The model building time of PLS runs in a fraction of the time compared to standard PCA and it results in similarly high precision models under the same circumstances as well as using only portions of the data set as training. Although there are newer supervised techniques than PCA, the precision problems presented by the technique are diminished when more latent dimensions are used, such as it is the case here. All tests favored PLS in terms of precision, followed by PCA. The others, although faster, presented worse precision.

As future work, we intend to employ incremental versions of PLS to provide fast changes of the model as needed when the ground truth changes.

Acknowledgements

The authors wish to acknowledge Brazilian financial agencies CNPq and FAPESP. We are also thankful to Frizzi San Roman for the preparation of the NEWS data set.

References

[BBKY06] BRONSTEIN M., BRONSTEIN A., KIMMEL R., YAVNEH I.: Multigrid Multidimensional Scaling. *Numerical Linear Algebra with Applications 13* (2006), 149–171. 2

[BGJ*97] BROADHURST D., GOODACRE R., JONES A., ROWLAND J., KELL D.: Genetic Algorithms as a Method for Variable Selection in Multiple Linear Regression and Partial Least Squares Regression, with Applications to Pyrolysis Mass Spectrometry. *Analytica Chimica Acta 348*, 1–3 (1997), 71–86. 3

[BN03] BELKIN M., NIYOGI P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput. 15*, 6 (2003), 1373–1396. 2

[BP07] BRANDES U., PICH C.: Eigensolver Methods for Progressive Multidimensional Scaling of Large Data. In *LNCS*, vol. 4372. Springer, 2007, pp. 42–53. 2

[BS07] BOULESTEIX A.-L., STRIMMER K.: Partial Least Squares: A Versatile Tool for the Analysis of High-Dimensional Genomic Data. *Briefings in Bioinformatics 8*, 1 (2007), 32–44. 2, 3

[dST04] DE SILVA V., TENENBAUM J.: *Sparse Multidimensional Scaling using Landmark Points*. Tech. rep., Department of Mathematics, Stanford University, CA, USA, 2004. 2

[Eld04] ELDEN L.: Partial Least-Squares vs. Lanczos Bidiagonalization–I: Analysis of a Projection Method for Multiple Regression. *Computational Statistics & Data Analysis 46*, 1 (2004), 11–31. 2

[ELS*08] ESS A., LEIBE B., SCHINDLER K., , VAN GOOL L.: A Mobile Vision System for Robust Multi-Person Tracking. In *IEEE CVPR* (Anchorage, AK, USA, June 2008), pp. 1–8. 5

[Gar94] GARTHWAITE P.: An Interpretation of Partial Least Squares. *Journal of the American Statistical Association 89*, 425 (1994), 122–127. 3

[Gel88] GELADI P.: Notes on the History and Nature of Partial Least Squares (PLS) Modelling. *Journal of Chemometrics 2*, 4 (1988), 231–246. 3

[HGP99] HOFFMAN P., GRINSTEIN G., PINKNEY D.: Dimensional Anchors: A Graphic Primitive for Multidimensional Multivariate Information Visualizations. In *Workshop on New Paradigms in Inform. Vis. and Manip. in Conjunction with ACM CIKM* (Kansas City, MO, USA, 1999), pp. 9–16. 2, 4

[JCC*11] JOIA P., COIMBRA D., CUMINATO J., PAULOVICH F., NONATO L.: Local Affine Multidimensional Projection. *IEEE TVCG 17*, 12 (2011), 2563–2571. 2

[Jol02] JOLLIFFE I.: *Principal Component Analysis*. Springer, 2002. 2

[KCH02] KOREN Y., CARMEL L., HAREL D.: ACE: A Fast Multiscale Eigenvectors Computation for Drawing Huge Graphs. In *IEEE Symp. on Inform. Visualiz.* (2002), pp. 137–144. 2

[KHD11] KEMBHAVI A., HARWOOD D., DAVIS L.: Vehicle Detection Using Partial Least Squares. *IEEE TPAMI 33*, 6 (2011), 1250–1265. 3

[Kru64] KRUSKAL J.: Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika 29* (1964), 115–129. 2

[LGB*95] LINDGREN F., GELADI P., BERGLUND A., SJOSTROM M., WOLD S.: Interactive Variable Selection (IVS) for PLS. Part II: Chemical Applications. *Journal of Chemometrics 9*, 5 (1995), 331–342. 3

[LN06] LEE S.-J., NOBLE A.: Use of Partial Least Squares Regression and Multidimensional Scaling on Aroma Models of California Chardonnay Wines. *American J. Enology and Viticulture 57*, 3 (Sept. 2006), 363–370. 2

[LW03] LI J., WANG J. Z.: Automatic Linguistic Indexing of Pictures by a Statistical Modelin Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25* (2003), 1075–1088. 5

[NOM*02] NESTOR P., O'DONNELL B., MCCARLEY R., NIZNIKIEWICZ M., BARNARD J., SHEN Z., BOOKSTEIN F., SHENTON M.: A New Statistical Method for Testing Hypotheses of Neuropsychological/MRI Relationships in Schizophrenia: Partial Least Squares Analysis. *Schizophrenia Research 53*, 1–2 (2002), 57–66. 3

[NR02] NGUYEN D., ROCKE D.: Tumor Classification by Partial Least Squares Using Microarray Gene Expression Data. *Bioinformatics 18*, 1 (2002), 39–50. 3

[PdRDK99] PEKALSKA E., DE RIDDER D., DUIN R., KRAAIJVELD M.: A New Method of Generalizing Sammon Mapping with Application to Algorithm Speed-up. In *Annual Conf. Advanced School for Comput. Imag.* (1999), pp. 221–228. 2

[PEP*11] PAULOVICH F., ELER D., POCO J., BOTHA C., MINGHIM R., NONATO L.: Piecewise Laplacian-based Projection for Interactive Data Exploration and Organization. *IEEE CGF, Proc. Eurovis 2011 30*, 3 (2011), 1091–1100. 2

[PFCP*11] PAIVA J., FLORIAN-CRUZ L., PEDRINI H., TELLES G., MINGHIM R.: Improved Similarity Trees and their Application to Visual Data Classification. *IEEE TVCG 17*, 12 (Dec. 2011), 2459–2468. 2

[PNML08] PAULOVICH F., NONATO L., MINGHIM R., LEVKOWITZ H.: Least Square Projection: A Fast High-Precision Multidimensional Projection Technique and Its Application to Document Mapping. *IEEE TVCG 14*, 3 (2008), 564–575. 2

[RS00] ROWEIS S., SAUL L.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science 290*, 5500 (Dec. 2000), 2323–2326. 2

[SGCD12] SCHWARTZ W. R., GUO H., CHOI J., DAVIS L. S.: Face Identification Using Large Feature Sets. *IEEE Transactions on Image Processing 21*, 4 (2012), 2245–2255. 3

[SKHD09] SCHWARTZ W., KEMBHAVI A., HARWOOD D., DAVIS L.: Human Detection Using Partial Least Squares Analysis. In *IEEE ICCV* (2009), pp. 24–31. 3

[TdSL00] TENENBAUM J., DE SILVA V., LANGFORD J.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science 290*, 5500 (Dec. 2000), 2319–2323. 2

[TLKT09] TALBOT J., LEE B., KAPOOR A., TAN D.: EnsembleMatrix: Interactive Visualization to Support Machine Learning with Multiple Classifiers. In *ACM CHI* (2009), pp. 1283–1292. 2

[Tor65] TORGESON W.: Multidimensional Scaling of Similarity. *Psychometrika 30* (1965), 379–393. 2

[TSK05] TAN P.-N., STEINBACH M., KUMAR V.: *Introduction to Data Mining*. Addison-Wesley Longman, Boston, MA, USA, 2005. 6

[Wol85] WOLD H.: Partial Least Squares. In *Encyclopedia of Statistical Sciences*, vol. 6. Wiley, New York, NY, USA, 1985, pp. 581–591. 1, 2, 3

[WPW03] WANG J., PENG W., WARD M.: Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration of High Dimensional Datasets. In *IEEE Symp. on Inform. Visualiz.* (Seattle, WA, USA, Oct. 2003), pp. 105–112. 2

[YWRH03] YANG J., WARD M., RUNDENSTEINER E., HUANG S.: Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets . In *Joint Eurographics - IEEE TVCG Symp. on Visualization* (Grenoble, France, 2003), pp. 19–28. 2