Stability Based Sparse LSI/PCA: Incorporating Feature Selection in LSI and PCA*

Dimitrios Mavroeidis[1] and Michalis Vazirgiannis[1,2]

[1] Department of Informatics, Athens University of Economics and Business, Greece
[2] GEMO Team, INRIA/FUTURS, France

Abstract. The stability of sample based algorithms is a concept commonly used for parameter tuning and validity assessment. In this paper we focus on two well studied algorithms, LSI and PCA, and propose a feature selection process that provably guarantees the stability of their outputs. The feature selection process is performed such that the level of (statistical) accuracy of the LSI/PCA input matrices is adequate for computing meaningful (stable) eigenvectors. The feature selection process "sparsifies" LSI/PCA, resulting in the projection of the instances on the eigenvectors of a principal submatrix of the original input matrix, thus producing sparse factor loadings that are linear combinations solely of the selected features. We utilize bootstrapping confidence intervals for assessing the statistical accuracy of the input sample matrices, and matrix perturbation theory in order to relate the statistical accuracy to the stability of eigenvectors. Experiments on several UCI-datasets verify empirically our approach.

1 Introduction

The intuitiveness of requiring that small changes in the input do not significantly affect the output of sample-based algorithms has made stability a very popular tool in machine learning. Many researchers have proposed the use of stability for assessing the validity (such as [14]) and for tuning the parameters of clustering algorithms (such as [17,13]). In this context an issue that presents several interesting challenges, is the analysis of the contribution that individual features have to the instability of the output and the derivation of necessary conditions that would guarantee stability, when a subset of the features is used. This analysis would allow for the introduction of feature selection algorithms that guarantee the stability of the output.

In this paper we focus on the stability of two well studied data preprocessing algorithms, Latent Semantic Indexing (LSI) [9], and Principal Components Analysis (PCA) [11]. These algorithms have been extensively used/studied in several machine learning papers (such as in [6,15]). Although they are not learning algorithms themselves, when they are used as a preprocessing step of a deterministic learning algorithm, their stability guarantees the stability of the output of the learning algorithm. Apart from

* This research project was co-financed by E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%). In particular, Dr. Vazirgiannis was supported by the Marie Curie Intra-European Fellowship NGWeMiS: Next Generation Web Mining and Searching (MEIF-CT-2005-011549).

J.N. Kok et al. (Eds.): ECML 2007, LNAI 4701, pp. 226–237, 2007.

the stability requirement, a motivation for performing feature selection stems from the fact that the factor loadings that are derived by LSI and PCA are linear combinations of all the input features, thus incorporating noise and making results difficult to interpret. The introduction of a feature selection process would resolve this issue, as it would result in sparse factor loadings that are linear combinations only of the selected features. Using analogous motivations, researchers have introduced "sparse" versions for various factor analysis techniques (an account of the related work can be found in section 3).

In general, the concept of stability is concerned with sample-based algorithms. In this setting, it is assumed that there exists a fixed (unknown) probability distribution that generates the data, and that the training set presents an i.i.d. sample drawn from this distribution. In the case of PCA and LSI, the input i.i.d. training set is used to construct the sample input matrices (the sample Covariance matrix in the case of PCA), that are essentially statistical estimates of the "true" input matrices that would be derived if we had complete knowledge of the data generating distribution. In this paper we focus on the instability that is related to the sampling variability (statistical accuracy) of the LSI/PCA input matrices, that can be naturally quantified using confidence intervals. In the context of our work, we assume that we do not have access to the data distribution and we utilize bootstrapping confidence intervals [10], which present a standard approach for measuring statistical accuracy (sampling variability) without making distributional assumptions.

In standard LSI/PCA the eigenvectors rely on the correlations/covariances between all the input features. However, some feature-correlations could be inaccurate in the statistical sense, thus degrading the quality of the resulting eigenvectors. In the proposed Stability based Sparse LSI/PCA (*SbS-LSI*, *SbS-PCA*) approach, we select the subset of features such that the level of accuracy of the term-term similarities/covariances is adequate for computing meaningful (stable) eigenvectors. Naturally this raises the need for determining the level of statistical accuracy that is needed for producing reliable (stable) eigenvectors. In order to address this issue we utilize matrix perturbation theory [20], which relates the eigenvalues and eigenvectors of matrices A and $A + E$. In order to employ matrix perturbation theory we consider A to be our input term-term similarity/covariance matrix and E to express the statistical inaccuracy of the term-term similarities/covariances, as quantified by the length of the respective bootstrap confidence intervals. *SbS-LSI* and *SbS-PCA* select a subset of the original features such that the eigenvectors of A' (the term-similarity/covariance matrix that is defined by the selected features, which is a principal submatrix of the original feature-similarity/covariance matrix) are stable with respect to perturbation E' (the respective principal submatrix of E).

The main innovation of *SbS-LSI* and *SbS-PCA* and the main differentiation from the related Sparse factor analysis approaches, is the fact that we utilize the stability criterion for "sparsifying" LSI/PCA (as we require the eigenvectors to be stable with respect to resampling variability). Concerning the practical impact of our work, experimental results on several real world UCI-datasets verify that the *SbS-LSI* and *SbS-PCA* algorithms can retrieve stable principal submatrices for various termination criteria.

2 Preliminaries

2.1 Latent Semantic Indexing and Principal Component Analysis

The Vector Space Model, used traditionally for representing documents, assumes that the terms are orthogonal, thus ignoring possible term-correlations. LSI aims at addressing this issue by projecting the documents to the k left singular vectors that corresponds to the k largest singular values of the term-document (feature-instance) Singular Value Decomposition (SVD). The SVD of a matrix A is defined as $A = U\Sigma V^T$ where U contains the left-singular vectors, V contains the right singular vectors and Σ contains the singular values. The left-singular matrix U can be considered as the eigenvector matrix of AA^T (the term-term similarity matrix). LSI projects the data using the equation: $A_k^T = V_k \Sigma_k$, or equivalently $A_k = U_k^T A$, where A_k is the new term-document matrix, containing only k dimensions (rows). There exist several variations of LSI, that have small differences with the generic approach described above. In this paper we adopt the variation where AA^T, the term similarity matrix, is derived by the cosine similarity measure rather than the inner-product.

Principal Components Analysis (PCA) is a dimensionality reduction technique that aims in retaining the maximal amount of variance in the projected space. PCA works by projecting the data in the first k eigenvectors (also called principal components) that correspond to the largest eigenvalues of the feature Covariance matrix. Thus, it can be observed that the low dimensionality transformations are derived by the same formula, as in LSI $A_k = U_k^T A$, with the difference being that the eigenvectors contained in U_k are derived by the feature Covariance matrix, instead of the feature-feature similarity (inner-product or cosine) matrix. This observation allows us to treat LSI and PCA in a uniform manner and define a feature selection framework that applies to both.

2.2 Bootstrapping

Bootstrapping [10] is a statistical method that can be used for measuring the accuracy of statistical estimates. In order to employ bootstrapping in estimating confidence intervals for the feature-feature similarities/covariances, we consider that the features are random variables with an unknown probability distribution. Thus, our input data can be considered as a random i.i.d. sample, where the observed feature values are derived by the unknown probability distribution. Taking the above into account, the cosine similarity/covariance abides to the definition of a statistic and as such its accuracy can be measured in the statistical sense.

Percentile Intervals present the most natural approach for constructing bootstrap confidence intervals. The Bias Corrected and Accelerated (BCa) bootstrap intervals present an improvement over the simple percentile intervals, in the sense that they account for the bias and the acceleration of the estimated parameter. The BCa intervals have several theoretical advantages over standard percentile intervals [10] that make them more appropriate in the context of our work.

A natural question that arises when employing the bootstrapping approach, for estimating confidence intervals, concerns the number of bootstrap samples needed to achieve accurate intervals. Based mainly on empirical evidence several researchers [10]

have reported that 1000 bootstrap samples are enough for accurately estimating bootstrap confidence intervals.

A more principled approach is presented in [1], where the authors introduce a three-step method, that allows for computing the number of bootstrap samples needed, in order to achieve a guaranteed accuracy with high probability. The level of approximation is user-defined and involves two parameters, the percentage deviation pdb, which measures the deviation of the computed interval from the ideal interval (i.e. the interval that is computed using infinite bootstrap samples) and the confidence τ. Using these parameters, their method computes the number of bootstrap samples that are sufficient for achieving the desired level of accuracy with probability $1 - \tau$.

2.3 Matrix Perturbation Theory

Matrix perturbation theory and more precisely Stewart's theorem on the perturbation of Invariant Subspaces [20] provides the means for assessing whether there exists a space that is spanned by k eigenvectors of the input term-term similarities/covariances that is not severely affected by the resampling variability (that quantifies the level of statistical inaccuracy) of the term-term similarities/covariances.

Although it would be more intuitive to consider the stability of the eigenvectors and not the spaces spanned by the eigenvectors, this would result in a not well-defined problem, since the eigenvectors that correspond to a tight cluster of eigenvalues are ill conditioned [20]. The inappropriateness of using eigenvectors can be also observed if we consider a matrix with two eigenvectors that have equal eigenvalues. In this case, it can be easily verified that any two orthogonal vectors in the subspace spanned by these two eigenvectors can be used to represent the eigenvectors of the original matrix. Theorem 1 presents a slightly modified version of the original Stewart's theorem, as presented by Papadimitriou et al. in [18]:

Theorem 1 (Stewart's theorem [20]). *Let A and $A + E$ be $n \times n$ symmetric matrices and let $V = [V_1\ V_2]$ be an orthogonal matrix, with $V_1 \in \mathbb{R}^{d \times n}$ and $V_2 \in \mathbb{R}^{(n-d) \times n}$, where $range(V_1)$ is an invariant subspace for A. Partition the matrices $V^T AV$ and $V^T EV$ as follows:*

$$V^T AV = \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix}$$

$$V^T EV = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$$

if

$$\delta = \lambda_{min} - \mu_{max} - ||E_{11}||_2 - ||E_{22}||_2 > 0$$

where λ_{min} is the smallest eigenvalue of Q_1 and μ_{max} is the largest eigenvalue of Q_2 and $||E_{12}||_2 \leq \delta/2$, then there exists a matrix $P \in \mathbb{R}^{(n-d) \times d}$ with $||P||_2 \leq \frac{2}{\delta}||E_{21}||_2$, such that the columns of $V_1' = (V_1 + V_2 P)(I + P^T P)^{\frac{1}{2}}$ form an orthonormal space that is invariant for $A + E$. Moreover, then

$$dist(range(V_1), range(V_1')) \leq \frac{2}{\delta}||E_{21}||_2$$

Using some elementary linear algebra we can simplify the above theorem and state that the space spanned by k eigenvectors that correspond to the largest k eigenvalues of matrix A will have small distance with the space spanned by k eigenvectors of matrix $A + E$, if the difference between the k-th and the $(k + 1)$-th eigenvalue of matrix A is at least 4 times larger than the Euclidean norm of the perturbation E.

2.4 Cauchy's Interlacing Theorem

Stewart's theorem, presented in the previous section, derives the stability of a matrix's eigenvector spaces by examining its eigenvalues. Thus, the naive approach for examining the stability of the matrix's principal submatrices, would be to compute all the respective eigenvalue decompositions. In order to avoid the prohibitive cost of computing all decompositions, we need to relate the eigenvalues of a matrix with the eigenvalues of its principal submatrices.

A well known theorem in the field of Linear Algebra that relates the eigenvalues of a matrix with the eigenvalues of its principal submatrices is Cauchy's Interlacing Theorem [20]. Cauchy's interlacing theorem is stated formally as follows:

Theorem 2 (Cauchy's Interlacing Theorem). *Let A be a matrix of order n with eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ and let B be a principal submatrix of A of order $n - 1$ with eigenvalues $\mu_1 \geq \mu_2 \geq ... \geq \mu_{n-1}$. Then $\lambda_1 \geq \mu_1 \geq \lambda_2 \geq \mu_2 \geq ... \geq \mu_{n-1} \geq \lambda_n$.*

Cauchy's Interlacing can be extended easily to rank $n - k$ principal submatrices.

Corollary 1. *Let A be a matrix of order n with eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ and let B be a principal submatrix of order $n-k$ of A with eigenvalues $\mu_1 \geq \mu_2 \geq ... \geq \mu_{n-k}$. Then $\lambda_i \geq \mu_i \geq \lambda_{i+k}$, $i = 1, 2, ..., n - k$.*

3 Related Work

In the context of supervised learning, the connection between the stability and the generalization performance has been studied (i.e in [4]). These studies derive generalization bounds for the performance of learning algorithms, based on their stability. In unsupervised learning the stability criterion has been used widely for choosing the parameters of clustering algorithms (i.e. selecting the number of clusters [17,13]), and for assessing the validity of clustering results [14].

Albeit the popularity of using stability for choosing the appropriate number of clusters, recent theoretical results have suggested that the stability criterion may not be appropriate for this task. More precisely, in [2] the authors prove for centroid based and spectral clustering that stability is determined by the symmetries of the data, which are not necessarily related to the parameters of the clustering algorithm. In [2] the authors focus on the stability concerning the structure of the clustering space and do not study the instability related to the sampling variability (which is related to the statistical accuracy). Our approach for "sparsifying" LSI/PCA is based on the analysis of the sampling variability affect, thus the result in [2] do not affect our approach.

The notion of stability has been studied within the contents of Principal Components Analysis (PCA). [3,8] present some early attempts to study stability of PCA by means of resampling (bootstrapping and jackknife). Potential applications of the stability criterion for PCA are presented in [3] where the stability is used for determining the appropriate dimension in PCA. Using different methodological approaches, the stability of the PCA has been also studied in [19].

Researchers working on PCA, have also identified the possible drawbacks of producing factor loadings that are a linear combination of all the input variables. This observation has led to attempts for "sparsifying" PCA [16,5,7,21]. Most of these approaches define the PCA problem as a cardinality constrained optimization problem and propose approximate algorithms for solving it. There are also a some research effort that use several heuristics for "sparsifying" LSI (such as [12]). Although we share the same motivation with Sparse PCA/LSI approaches, we utilize the stability criterion for "sparsifying" LSI/PCA. This differentiates significantly our approach from the Sparse PCA/LSI approaches.

4 Stability Based Sparse LSI/PCA

After introducing all the necessary notions and having presented the related work, we can move on to describe our proposed approach for "sparsifying" LSI/PCA. Recall that the main intuition in the proposed approach for Sparse LSI/PCA, is to select a feature subset such that the level of accuracy of the term-term similarities/covariances is adequate for computing stable PCA/LSI solutions.

4.1 Measuring Resampling Variability of Term-Term Similarities/Covariances

For quantifying the resampling variability of the estimated term-term similarities/ covariance, we utilize a principled statistical approach, bootstrapping BCa confidence intervals, that enable the non-parametric calculation of the statistical accuracy of the sample term-term similarities/covariances. We illustrate the method adopted, with the following example:

Example 1. Consider that we have the terms *"graduation"* and *"unemployment"* that are contained in a sample of 5 documents. The sample can be represented as a set of 2-tuples, where each 2-tuple contains the frequency of occurrence of the two terms in each document $\{d_1, d_2, d_3, d_4, d_5\} = \{(0,1),(2,3),(4,3),(3,0),(1,0)\}$. The cosine similarity of the two terms is 0.75. Taking only the cosine into account one could derive that the two terms are semantically similar, however if we take 5 bootstrap samples we may have:

bootstrap sample	cosine
{(0,1),(0,1),(4,3),(3,0),(1,0)}	0.70
{(0,1),(2,3),(3,0),(3,0),(1,0)}	0.40
{(0,1),(2,3),(4,3),(2,3),(1,0)}	0.96
{(0,1),(3,0),(1,0),(3,0),(1,0)}	0.0
{(2,3),(2,3),(4,3),(2,3),(4,3)}	1.0

The percentile confidence interval at 0.6 coverage is [0.0,0.96], which quantifies the variability of the values of the cosine similarity with respect to resampling of the input data. This implies that more data are needed in order to conclude on the two terms semantic similarity.

In the case we want to compute the confidence interval for the covariance, we can work in an analogous manner, using the covariance instead of the cosine formula. From the example it can be observed that the calculated interval may have variability itself (different runs could result in different intervals). In order to address this issue, we use 1000 bootstrap samples for calculating the confidence intervals. As it has been mentioned in the preliminaries section, 1000 bootstrap samples are considered to be adequate for computing accurate confidence intervals [10].

4.2 Relating Variability to Stable Sub-spaces

Since we aim at selecting stable sub-spaces with respect to the resampling variability of the term-term similarities/covariances, we need the means for relating the variability expressed by the BCa intervals (which quantify the statistical accuracy of the term-similarities/covariances) to the stability of the eigenvectors of the term-term similarity/covriance matrix. For this purpose we utilize Matrix Perturbation Theory and more precisely Stewart's Theorem that essentially relates the eigenvector spaces of matrices A and $A+E$, with respect to the eigenvalues of A and the norm of E. In the context of our work A contains the term-term similarities/covariances, while $A+E$ contains the perturbed version of A.

For determining the elements of matrix E we adopt the conservative approach of computing the elements of matrix E as the maximum difference between the term-term similarities/covariances and the endpoints of the corresponding confidence intervals. The intuition behind selecting the matrix E to contain the largest differences stems from a property of the Euclidean norm stating that $||E||_2$ lies in the interval $[\frac{1}{\sqrt{n}}||E||_1, \sqrt{n}||E||_1]$ where n is the number of columns (or rows) of E and $||E||_1 = max_{1\leq j\leq n}\sum_i^n |a_{ij}|$ (a_{ij} are the elements of the matrix). Thus by defining the elements of matrix E to include the maximum differences, we force the $||E||_2$ to lie within an interval of larger values (worst case scenario). Moreover this property of the Euclidean norm justifies the heuristic of removing the rows and columns with highest norm (later introduced in Algorithm 1). Since $||E||_1$ is defined using the absolute values of the matrix elements, it makes no difference (in the estimated interval) if we include negative values in the definition of E. In order to illustrate our approach, we provide the following example (continuation of example in section 4.1):

Example 2. Consider the case where we have terms i: *"graduation"* and j: *"unemployment"*, then taking after the example in section 4.1 we will have the $(i,j)^{th}$ and $(j,i)^{th}$ element of matrix A to be $A(i,j) = 0.75$. Since the confidence interval produced by bootstrap samples is [0.0,0.96], the $(i,j)^{th}$ and $(j,i)^{th}$ element of matrix E will be determined by $E(i,j) = max\{|0.75-0.0|, |0.75-0.96|\} = 0.75$.

Having determined the elements of matrix E we can apply Stewart's theorem in a straight forward manner. This will allow for assessing whether there exists some k for

which the space spanned by the k eigenvectors is stable. If no such k exists, we should investigate sparser representations, removing the terms that exhibit high variance in their similarity estimations. The problem of searching for stable principal submatrices presents several challenges which we investigate in the subsequent section.

4.3 Stable Principal Submatrices

As we have argued in the introductory section, the problem of identifying stable principal submatrices is related to the problem of feature selection for LSI/PCA. In our approach we formulate two requirements for selecting the feature subset.

1. The feature-feature similarity/covariance matrix induced by the subset of features (which is a principal submatrix of the original similarity/covariance matrix), should contain stable eigenvector spaces for some k.
2. The E matrix induced by the subset of features (which is a principal submatrix of the original E matrix) should contain the most accurate similarity/covariance estimations.

The main difficulty for identifying stable principal submatrices is that the stability of the eigenvector spaces is assessed using the eigenvalues of the matrix. Thus, the naive approach would require that the eigenvalue decomposition is performed on every candidate matrix, making the cost of searching stable submatrices prohibitive.

In order to reduce the computational cost, we make use of Proposition 1, that is derived from the Cauchy's interlacing theorem. Proposition 1 allows for evaluating, whether it is possible for stable eigenvector spaces to exist in a principal submatrix, prior to computing its eigenvalue decomposition. Consequently the number of eigendecompositions that are performed can be significantly reduced. In Proposition 1, we do not check directly the requirements of Stewart's theorem but the requirements stated in Lemma 1. The proofs for Lemma 1 and Proposition 1 can be found in the Appendix.

Lemma 1. *Let A and $A+E$ be $n \times n$ symmetric matrices and let the eigenvalues of A and E be $\lambda_1^{(A)} \geq ... \geq \lambda_n^{(A)}$ and $\lambda_1^{(E)} \geq ... \geq \lambda_n^{(E)}$ respectively. If $\lambda_1^{(E)} > 0$ and $\lambda_i^{(A)} - \lambda_{i+1}^{(A)} > 4 \cdot \lambda_1^{(E)}$ for some i, then the prerequisites of Stewart's theorem will hold, and the space spanned by the first i eigenvectors of A will be stable.*

Proposition 1. *Let A and E be matrices of order n with eigenvalues $\lambda_1^{(A)} \geq \lambda_2^{(A)} \geq ... \geq \lambda_n^{(A)}$ and $\lambda_1^{(E)} \geq \lambda_2^{(E)} \geq ... \geq \lambda_n^{(E)}$ respectively. Moreover let A' and E' be principal submatrices of A and E of order $n-k$. If $\lambda_i^{(A)} - \lambda_{i+k+1}^{(A)} \leq 4 \cdot \lambda_{1+k}^{(E)}$, for all $i = 1, 2, ..., n-k-1$, then the prerequisites of lemma 1 do not hold for matrices A' and $A'+E'$.*

In order to retrieve stable principal submatrices, we adopt the approach of incrementally reducing the order of the input term-term similarity/covariance A_n matrix at each step by 1. The choice of the principal submatrix of order $n-1$ is done with respect to retaining the most accurate term-term similarities/covariances. This is done by removing the term that corresponds to the row (and column) of E_n that has the highest norm. Subsequently, using Proposition 1 we check whether it is possible for A_{n-1} to

satisfy the prerequisites of Lemma 1 (and thus to contain stable eigenvector spaces). Until this condition is satisfied, we continue to reduce the order of the similarity matrix by 1. When proposition 1 cannot guarantee that the prerequisites of Lemma 1 will fail, we compute the eigenvalue decomposition and verify analytically whether the prerequisites of Lemma 1 hold. If we derive (using Lemma 1) that there exist stable eigenvector spaces, then we can check the standard LSI/PCA termination criteria (i.e. in PCA we can require that the stable eigenvector spaces retain a required amount of the initial variance). If the standard termination criteria are met, then we terminate the algorithm and output the stable principal submatrix. Otherwise we continue to iterate until such such stable submatrix is found. Our algorithm for detecting stable principal submatrices is illustrated as Algorithm 1.

Algorithm 1. SbS-LSI/PCA(S,A,E,TerminationCriteria)

```
Compute the eigenvalues of A
Compute the eigenvalues of E
if (The prerequisites of Lemma 1 are satisfied) AND (TerminationCriteria are met) then
    return A and the k for which the TerminationCriteria are met.
else
    repeat
        Find row r (feature t) of E with the highest norm
        S' ←Remove from S feature t
        A' ←Remove from A row and column r
        E' ←Remove from E row and column r
    until Proposition 1 cannot guarantee that the preconditions of Lemma 1 will not be satisfied
    call SbS-LSI/PCA(S',A',E',TerminationCriteria)
end if
```

In order to illustrate the use of the traditional termination criteria, consider that we are provided with a dataset that contains m objects with n features. This input dataset induces an initial $n \times n$ covariance matrix C. Moreover, consider that we aim in projecting the data in a stable sub-space such that $a\%$ of the variance of the original input data is retained. Each time we find a principal submatrix of C that has stable eigenvectors, it will contain stable eigenspaces for a certain number k of eigenvectors (i.e. $k = k_1$,$k = k_2$ and $k = k_3$, recall that stability for a certain number of eigenvectors k depends on the difference of eigenvalues λ_k and λ_{k+1}, thus it can be achieved for various values of k). Then we should check whether retaining k (for all possible "stable values" of k) eigenvectors of the principal submatrix is adequate for satisfying the termination criterion set (i.e. expressing $a\%$ of the original data variance).

5 Experiments

In the experimental section we empirically demonstrate the behavior of *SbS-LSI* in three real world UCI-datasets. The datasets are the Ionosphere dataset that contains radar data (with 351 objects, 35 dimensions and 2 class labels), the Segmentation dataset that

contains outdoor image segmentations (with 2000 objects, 19 dimensions and 7 class labels) and the Spambase dataset that contains emails classified as spam/non spam (with 4601 objects, 58 attributes and 2 class labels). In the experimental setup, we have set the confidence level to 0.9 for the BCa intervals. Moreover, we consider that the termination criteria for *SbS-LSI* is set to be a certain proportion of the Frobenius norm of the original data (i.e. we require that the projected space expresses $a\%$ of the Frobenius norm of the original space). Usually, in the context of LSI, the termination criterion is set to be a number k that represents the number of eigenvectors used to project the data (and thus the dimensionality of the reduced space), however this is not appropriate in *SbS-LSI* as the k values have different semantics with respect to different principal submatrices (the first k eigenvectors of a 500×500 matrix are not comparable to the first k eigenvectors of a 5000×5000 matrix).

In the Ionosphere dataset, we can easily evaluate using Proposition 1 that the input data (using all the features), do not contain stable eigenvectors. If we set the proportion $a \geq 33\%$, then the algorithm removes 2 features that contribute maximally to the instability of LSI and finds a principal submatrix that has stable eigenvector spaces for $k = 1$. Using the values of the respective eigenvectors, the algorithm verifies that the termination criteria (concerning the proportion a) is satisfied and the algorithm terminates. A similar behavior is exhibited by the Spambase dataset, that does not contain stable eigenvectors in the original space. If we set the proportion $a \geq 25\%$, then the algorithm removes 43 features and finds a principal submatrix that has stable eigenvector spaces for $k = 13$. Then the algorithm verifies that the termination criteria (concerning the proportion a) is satisfied and thus the algorithm terminates.

A more interesting behavior is exhibited by the Segmentation datasets. If we set the proportion $a \geq 45\%$, then the algorithm, verifies that the first eigenvector of the original input matrix (using all the features) is stable and satisfies the termination criteria. However, if we set $a \geq 55\%$, then the input matrix is not adequate and *SbS-LSI* has to examine the stability of principal submatrices. For $a \geq 55\%$ the algorithm removes 3 features and finds a principal submatrix that has stable eigenvector spaces for $k = 1$ and $k = 3$. For $k = 3$ the termination criteria are met, and the algorithm terminates. If we set $a \geq 65\%$, then the algorithm removes another two features and retrieves a principal submatrix that has stable eigenvectors for $k = 1, 2, 3, 4$. For $k = 4$, the termination criteria are met and the algorithm terminates.

The experiments demonstrate the practical applications of *SbS-LSI* and *SbS-PCA*. If we take into account the intuitiveness of the stability criterion, then it is natural to consider preferable to choose the stable sub-spaces that are derived by *SbS-LSI* and *SbS-PCA* over the standard LSI and PCA results. For example in the Segmentation dataset if we set as a termination criteria $a \geq 65\%$, standard LSI will return a certain number of k eigenvectors from the original input data, that are not stable with respect to sampling variability. On the contrary *SbS-LSI* performs feature selection, removes 5 features that contribute maximally to the instability of LSI, and returns $k = 4$ eigenvector of the resulting principal submatrix that are guaranteed to be stable. In the cases where the termination criteria are not met (i.e. no stable principal submatrix is found) , then it would be a sound practice to avoid using LSI or PCA as a preprocessing step to machine learning algorithms.

6 Conclusions and Further Work

Motivated by the intuitiveness of the stability criterion, we have introduced a feature selection process for "sparsifying" LSI and PCA. The proposed *SbS-LSI* and *SbS-PCA* algorithms select a feature subset such that the level of the statistical accuracy of the term-term similarities/covariances is adequate for computing stable eigenvectors (and thus stable sub-spaces). The main theoretical innovation of *SbS-LSI* and *SbS-PCA* is the fact that they present the first approach that utilizes the concept of stability for "sparsifying" LSI/PCA.

Concerning further work, we aim at investigate possible solutions for reducing the computational cost of retrieving stable submatrices. We also intend to investigate the theoretical properties of using the stability criterion for sparsifying LSI/PCA and its potential applications in Spectral Clustering.

References

1. Andrews, D.W., Buchinsky, M.: On the number of bootstrap repetitions for BCa confidence intervals. Econometric Theory (2002)
2. Ben-David, S., von Luxburg, U., Pal, D.: A sober look at clustering stability. In: Lugosi, G., Simon, H.U. (eds.) COLT 2006. LNCS (LNAI), vol. 4005, Springer, Heidelberg (2006)
3. Besse, P.: PCA stability and choice of dimensionality. Statistic & Probability Letters (1992)
4. Bousquet, O., Elisseeff, A.: Stability and generalization. Journal of Machine Learning Research (2002)
5. Cadima, J., Jolliffe, I.T.: Loadings and correlations in the interpretation of principal components. Journal of Applied Statistics (1995)
6. Cristianini, N., Shawe-Taylor, J., Lodhi, H.: Latent Semantic Kernels. Journal of Intelligent Information Systems (2002)
7. d'Aspremont, A., Ghaoui, L.E., Jordan, M.I., Lanckriet, G.R.G.: A direct formulation for sparse PCA using semidefinite programming. In: NIPS (2004)
8. Daudin, J., Duby, C., Trecourt, P.: Stability of principal component analysis studied by the bootstrap method. Statistics (1988)
9. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society For Information Science (1990)
10. Efron, B., Tibshirani, R.: An introduction to the bootstrap. Chapman Hall (1993)
11. Jolliffe, I.T.: Principal Components Analysis. Springer, Heidelberg (2002)
12. Kontostathis, A., Pottenger, W.M., Davison, B.D.: Identification of critical values in Latent Semantic Indexing (LSI). In: Lin, T.Y., Ohsuga, S., Liau, C.J., Tsumoto, S. (eds.) Foundations of Data Mining and Knowledge Discovery, Springer, Heidelberg (2005)
13. Lange, T., Roth, V., Braun, M.L., Buhmann, J.M.: Stability-based validation of clustering solutions. Neural Computation (2004)
14. Levine, E., Domany, E.: Resampling method for unsupervised estimation of cluster validity. Neural Computation (2001)
15. Mika, S., Schölkopf, B., Smola, A.J., Müller, K.-R., Scholz, M., Rätsch, G.: Kernel PCA and De-Noising in Feature Spaces. In: NIPS 1998 (1998)
16. Moghaddam, B., Weiss, Y., Avidan, S.: Spectral bounds for sparse PCA: Exact and greedy algorithms. In: NIPS 2005 (2005)
17. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning (2003)

18. Papadimitriou, C.H., Raghavan, P., Tamaki, H., Vempala, S.: Latent semantic indexing: A probabilistic analysis. In: PODS 1998 (1998)
19. Shawe-Taylor, J., Williams, C.K.I.: The stability of kernel principal components analysis and its relation to the process eigenspectrum. In: NIPS 2002 (2002)
20. Stewart, G., Sun, J.-G.: Matrix perturbation theory. Academic Press, London (1990)
21. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. Journal of Computational and Graphical Statistics (2006)

Appendix

Proof (Lemma 1). We have:

$\lambda_i^{(A)} - \lambda_{i+1}^{(A)} > 4 \cdot \lambda_1^{(E)} \Rightarrow$
$\lambda_i^{(A)} - \lambda_{i+1}^{(A)} > 4 \cdot ||E||_2 \Rightarrow$
$\lambda_i^{(A)} - \lambda_{i+1}^{(A)} - 2 \cdot ||E||_2 > 2 \cdot ||E||_2 \Rightarrow$
$\lambda_i^{(A)} - \lambda_{i+1}^{(A)} - ||E_{11}||_2 - ||E_{22}||_2 \geq \lambda_i^{(A)} - \lambda_{i+1}^{(A)} - 2 \cdot ||E||_2 > 2 \cdot ||E||_2 \geq 2 \cdot ||E_{12}||_2$ (eq. 1)

From equation 1 we can derive $\delta \geq 2 \cdot ||E_{12}||_2$
Moreover we have:

$\lambda_1^{(E)} > 0 \Rightarrow ||E||_2 > 0 \Rightarrow$(using equation 1) $\delta > 0$

Thus the prerequisites set by Stewart's theorem are met and the space spanned by the first i eigenvectors of matrix A sill be stable under perturbation E.

Note that for deriving the result we have used the fact that $||E||_2 \geq ||E_{ij}||_2$ and the definition for δ from Stewart's theorem.

Proof (Proposition 1). Let the eigenvalues of A' and E' be $\mu_1^{(A')} \geq ... \geq \mu_{n-k}^{(A')}$ and $\mu_1^{(E')} \geq ... \geq \mu_{n-k}^{(E')}$ respectively.

Concerning the eigenvalues of A and A' we can derive from corollary 1 that:

$$\begin{matrix} \mu_i^{(A')} \leq \lambda_i^{(A)} \\ \mu_{i+1}^{(A')} \geq \lambda_{i+1+k}^{(A)} \end{matrix} \Rightarrow \mu_i^{(A')} - \mu_{i+1}^{(A')} \leq \lambda_i^{(A)} - \lambda_{i+1+k}^{(A)}$$

Concerning the eigenvalues of E and E' we can derive from corollary 1 that $\lambda_{1+k}^{(E)} \leq \mu_1^{(E')}$

Thus if we have:

$\lambda_i^{(A)} - \lambda_{i+k+1}^{(A)} \leq 4 \cdot \lambda_{1+k}^{(E)}$ for all $i = 1, 2, ..., n-k-1 \Rightarrow$
$\mu_i^{(A')} - \mu_{i+1}^{(A')} \leq 4 \cdot \mu_1^{(E')}$ for all $i = 1, 2, ..., n-k-1$

Thus the prerequisites of Lemma 1 do not hold for matrices A' and $A' + E'$.