A Novel Stability Based Feature Selection Framework for k-means Clustering*

Dimitrios Mavroeidis and Elena Marchiori

Institute for Computing and Information Sciences,
Radboud University Nijmegen, The Netherlands

Abstract. Stability of a learning algorithm with respect to small input perturbations is an important property, as it implies the derived models to be robust with respect to the presence of noisy features and/or data sample fluctuations. In this paper we explore the effect of stability optimization in the standard feature selection process for the continuous (PCA-based) k-means clustering problem. Interestingly, we derive that stability maximization naturally introduces a tradeoff between cluster separation and variance, leading to the selection of features that have a high cluster separation index that is not artificially inflated by the feature's variance. The proposed algorithmic setup is based on a Sparse PCA approach, that selects the features that maximize stability in a greedy fashion. In our study, we also analyze several properties of Sparse PCA relevant to stability that promote Sparse PCA as a viable feature selection mechanism for clustering. The practical relevance of the proposed method is demonstrated in the context of cancer research, where we consider the problem of detecting potential tumor biomarkers using microarray gene expression data. The application of our method to a leukemia dataset shows that the tradeoff between cluster separation and variance leads to the selection of features corresponding to important biomarker genes. Some of them have relative low variance and are not detected without the direct optimization of stability in Sparse PCA based k-means.

1 Introduction

The stability of a learning algorithm with respect to small input perturbations is generally considered a desired property of learning algorithms, as it ensures that the derived models are robust and are not significantly affected by noisy features or data sample fluctuations. Based on these motivations, the notion of stability has been employed by several popular machine learning paradigms (such as Bagging) and it has been the central theme in several studies that focus both on the theoretical study of stability and the development of practical stability optimizing algorithms. Albeit the considerable amount of research that has been devoted to the study of stability, the interplay between clustering stability and feature selection has not been substantially investigated. This is because most feature selection frameworks do not take into account the contribution of the features to the variance of the derived models and solely evaluate the "relevance" of each feature to the target class structure. This may result in suboptimal models since

* This work was partially supported by the Netherlands Organization for Scientific Research (NWO) within NWO project 612.066.927.

D. Gunopulos et al. (Eds.): ECML PKDD 2011, Part II, LNAI 6912, pp. 421–436, 2011.

prediction error is, as illustrated by the bias-variance decomposition, affected by both the relevance of each feature (bias) and its contribution to the stability (variance) of the resulting data model. These considerations, that are also discussed in [13] motivate the study for practical feature selection algorithms that achieve the right balance between the bias-variance tradeoff and optimize the predictive ability of the resulting models.

In the context of this work we undertake this challenge and explore the potentials of performing feature selection with the general purpose of maximizing the stability of the continuous (PCA-based) k-means clustering output[1]. The proposed analysis is performed at a theoretical, algorithmic and empirical level, which are summarized in the sequel. From the theoretical point of view, we demonstrate that stability maximization, naturally leads to a cluster separation vs. feature variance trade off that results in the selection of features that have a high cluster separation index that is not artificially inflated by the feature's variance. This conceptual contribution brings new insights to the theoretical properties of stability, provides practitioners with a clear understanding as to when the stability maximizing objective is appropriate in a specific application context and also allows for the effective interpretation of the success (or possible failure) of the stability based feature selection process.

From the algorithmic point of view, we propose a Sparse PCA formulation for selecting the relevant features that maximize the stability of the continuous clustering solution. Sparse PCA presents a natural choice, since the continuous k-means solution is derived by the principal components, i.e. the dominant eigenvectors of the feature covariance matrix [5]. In our study of Stable Sparse PCA we derive several interesting results that are related to the suitability of Sparse PCA for feature selection in clustering and also, to the stability of the Sparse PCA output. Specifically, we demonstrate that Sparse PCA can be derived as a continuous relaxation to a feature selection problem that optimizes for a cluster separation index. Moreover, we show that double centering the data before the application of Sparse PCA, leads to a "two-way" stability property. I.e. the stability of the instance-clusters becomes equal to the stability of the feature-clusters. This is an important observation that complements our work with data mining algorithms that utilize the feature clusters in the data mining process. Finally, we propose a novel "two-way" stable Sparse PCA algorithm that relies on a greedy lower bound optimization. These results can be considered as side-contributions to our understanding of Sparse PCA as a feature selection mechanism for clustering.

Empirically, we verify the proposed Stable Sparse PCA framework in the context of Cancer Research. In our experiments we have employed four publicly available microarray datasets that are related to the identification of certain cancer types. The experiments demonstrate that the proposed Stable Sparse PCA method is competitive and often superior to state-of-the-art feature selection methods. In particular, we consider the problem of detecting potential tumor biomarkers using microarray gene expression data. Application of our method to leukemia gene expression data shows that the tradeoff between cluster separation and variance leads to the selection of features corresponding to important biomarker genes. Some of them have relative low variance and are not detected without the direct optimization of stability.

[1] With the term continuous k-means clustering problem we refer to the continuous relaxation approach for approximating k-means [5].

2 Spectral k-means

K-means clustering is arguably the most popular clustering algorithm among data mining practitioners, and albeit its introduction more than 50 years ago, it still constitutes an active area of research. The goal of K-means is to find the clustering that minimizes the sum of squared distances of the elements of each cluster to the cluster centers. Formally this objective can be stated as: $J_K = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k||^2$ where we consider x_i to be the instance vectors, μ_k the respective cluster centers and C_k, to denote the clusters.

The most popular heuristic for approximating J_K is the standard Lloyd's algorithm, that starts with a random initial guess of the cluster center and iteratively converges using an EM-style iterative process to a local optima of the k-means objective. In the context of this work, we focus on a different approximation scheme for the k-means objective that is based on the continuous (spectral) relaxation of the discrete cluster assignment vector[5]. The Spectral relaxation allows us to study the stability of the clustering output using the advanced results of matrix perturbation theory [18].

In order to illustrate spectral k-means, we recall from [5] that the k-means problem can be written in equivalent form as: $J_K = \mathbf{Trace}(X_{fc}^T X_{fc}) - \frac{1}{2} J_D$ where X_{fc} is the input $m \times n$ feature-instance matrix, with centered features (rows), and J_D in the 2-cluster case (clusters c_1 and c_2 with sizes n_1 and n_2) is defined as:

$$J_D = \frac{n_1 n_2}{n} \left[2\frac{d(c_1,c_2)}{n_1 n_2} - \frac{d(c_1,c_1)}{n_1^2} - \frac{d(c_2,c_2)}{n_2^2} \right] \quad (1)$$

with $d(c_k,c_l) = \sum_{i \in C_k, j \in C_l} ||x_i - x_j||^2$. Moreover, in [5] it is demonstrated that $J_D = 2\mathbf{Trace}(Q_{K-1}^T X_{fc}^T X_{fc} Q_{K-1})$, where Q_{K-1} is a $n \times (K-1)$ matrix (n =#inst., K =#clust.) that contains the discrete cluster labels (for the discrete cluster values of matrix Q_{K-1} we refer to [5]).

Based on the afore equations, the minimization of J_K is equivalent to the maximization of J_D. By applying the continuous relaxation to J_D, the *continuous* solution is derived by projecting the data to the $k-1$ principal eigenvectors, i.e. the $k-1$ dominant eigenvectors of the Covariance matrix that correspond to the largest eigenvalues. Naturally, the spectral solution will contain the continuous values and an extra step needs to be applied to discretize the continuous cluster assignments with a popular heuristic being the application of standard Lloyd's k-means to the reduced principal eigenspace.

It can be noticed that the minimization of J_K is equivalent to the maximization of J_D because $\mathbf{Trace}(X_{fc}^T X_{fc})$ is a constant that is equal to the (scaled) sum of variances of the available features[2]. In a feature selection setup this term will not remain constant since different features may have different variances, unless the data are appropriately preprocessed such that they have equal variances.

The stability of Spectral k-means can be evaluated using Matrix Perturbation Theory [18]. The relevant theorems designate that the stability of the continuous Spectral k-means solution depends on the size of the eigengap $\lambda_{k-1} - \lambda_k$ between the $k-1$ and the k largest eigenvalues of the relevant matrix with a larger eigengap implying improved stability. Thus, a stability optimizing algorithm should aim to maximize this eigengap.

[2] This is because $\mathbf{Trace}(X_{fc}^T X_{fc}) = \mathbf{Trace}(X_{fc} X_{fc}^T)$.

3 Stable Sparse PCA

3.1 Stability Maximizing Objective and the Cluster Separation/Variance Tradeoff

We will now move on to define the appropriate optimization objective for feature selection that maximizes the stability of the Spectral k-means clustering output. The proposed formulation is based on the Sparse PCA approach in [3], with the appropriate modifications that account for stability maximization.

In order to optimize for stability, we incorporate a term that accounts for the difference between the two largest eigenvalues. Since the aim is to distinguish between the two largest eigenvalues and not between λ_{k-1} and λ_k, our framework initially considers the two-way clustering problem, and extends for $k > 2$-way clustering using a deflation method analytically described in Section 4.3. In this manner, the proposed framework can select a different subset of features at each sequential step (for each eigenvector) thus possibly identifying different feature subsets for separating between different
clusters.

For facilitating the optimization problem we consider the average difference between the largest eigenvalue with the rest. I.e. $\frac{1}{n}\sum_{i=1}^{n}(\lambda_1 - \lambda_i)$ instead of $\lambda_1 - \lambda_2$. Although this formulation will not directly optimize for the difference between the largest eigenvalues, the objective will have a stronger incentive for minimizing the eigenvalues that are closer to λ_1 since they will contribute more to the maximization of the average difference. As we will illustrate in this section, the difference between the largest and the consecutive eigenvalues gives rise to a tradeoff between maximizing the distance between clusters and the feature variances. This balancing essentially imposes a variance based threshold on a cluster separation index and utterly selects the features that optimize the harder separation objective.

Before we move on to define our objective function we will clarify the notation we will use. We will denote X as our input $m \times n$ (feature-instance) matrix, $C_n = (I - e_n e_n^T/n)$, denotes the standard row (feature) centering matrix, u is a vector of length m, with $u(i) = 1$ if feature i is retained in the final solution. $\mathbf{diag}(u)$ is an $m \times m$ diagonal matrix with vector u in its diagonal (i.e. $u(i,i) = 0$ if feature i is removed, otherwise it is equal to 1), and $\mathbf{card}(u)$ is equal to the number of non-zero elements of u (i.e. the number of features that are selected. It can be observed that the multiplication $\mathbf{diag}(u)X$ essentially removes the features that correspond to $u(i) = 0$. Finally we will denote the column (feature)-centered matrix as $X_{fc} = XC_n$ and also $x_{fc}(i)$ as a $n \times 1$ vector that contains the centered vector-representation of feature i (notice we represent $x_{fc}(i)$ as a column vector although it corresponds to row i of matrix X_{fc}).

Based on the afore notation, the covariance matrix after feature selection (omitting term $1/n$ that does not affect our optimization problem) is defined as:

$$Cov = \mathbf{diag}(u)X_{fc}X_{fc}^T\mathbf{diag}(u)$$

We now define the stability maximizing objective as:

$$\begin{aligned} Obj &= \max_{u\in\{0,1\}^m} \left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(\lambda_1(Cov)-\lambda_i(Cov))\right) \\ &= \max_{u\in\{0,1\}^m} \left(\tfrac{n-1}{n}\lambda_1(Cov)-\tfrac{1}{n}\textstyle\sum_{i=2}^{n}\lambda_i(Cov))\right) \\ &= \max_{u\in\{0,1\}^m} \left(\lambda_1(Cov)-\tfrac{1}{n}\mathbf{Trace}(Cov)\right) \end{aligned} \tag{2}$$

Based on the ability to express the k-means objective using the clustering separation index J_D (as analyzed in Section 2), we can derive the afore objective as a continuous relaxation to the following feature selection clustering objective:

$$\max_{u\in\{0,1\}^m} \frac{n_1 n_2}{n}\left[2\frac{d^{(u)}(c_1,c_2)}{n_1 n_2}-\frac{d^{(u)}(c_1,c_1)}{n_1^2}-\frac{d^{(u)}(c_2,c_2)}{n_2^2}\right]-\sum_{i=1}^{m}u_i\cdot \mathrm{var}(f_i) \tag{3}$$

where $d^{(u)}(c_i,c_j)=\sum_{k\in c_i}\sum_{l\in c_j}(x_k^{(u)}-x_l^{(u)})^2$ and $x^{(u)}$ denotes the representation of an instance after feature selection (i.e. only the selected features are taken into account when computing the respective distances). Moreover, $\mathrm{var}(f_i)$ denotes the variance of feature i. Notice that the cluster separation index is essentially the J_D of formula 1 after feature selection. The proof of the relationship between the Objectives 2,3 is mostly based on the derivations made within [5] and is omitted due to space limitations.

Based on the afore analysis, we have demonstrated that stability optimization leads to the introduction of a cluster separation vs. variance tradeoff in the feature selection process. In this manner the features that are selected will have high cluster separation value and among features with equal cluster separation value the ones with the smaller variance will be selected. The novelty of the proposed objective resides in the fact that it explicitly penalizes high feature variance and it leads to the selection of the feature subset that has high cluster separation index and low variance. Although this seems to contradict a basic rule of thumb in feature selection that considers features with high variance to be more helpful in separating between clusters (notably, the selection of features with high variance is commonly used as a baseline in the empirical evaluation of feature selection algorithms), our framework can be justified by the view of feature selection as a *variance reduction* process (as done in [13]). In this conceptual approach, feature selection improves the quality of a learning algorithm when it achieves the reduction of variance without significantly increasing the algorithm's bias. Interestingly, under this paradigm the contribution of each feature to the bias-variance of the output model is more important that the exact identification of the relevant/non-relevant features (i.e. a relevant feature that contributes highly to the model's variance may not be desirable).

The Stable Sparse PCA objective formulation currently accounts only for the maximization of the stability of the instance-clustering output. We use the term"solely" as we will demonstrate in the next section that this objective can be extended such that it simultaneously optimizes the stability of both instance and feature clusters.

3.2 Two-Way Stability

Several data mining frameworks employ the clustering of the features as an important component within the general data mining process. One such example is bi-clustering,

or co-clustering [4] where one tries to cluster simultaneously the features and the instances for identifying the clusters of features that can be used for describing the instance clusters. In these application contexts the stability of the clustering of the features is of central importance, since an unstable cluster structure could result in spurious feature clusters that are sensitive to noise or data sample variations.

Based on these motivations, we will present here the necessary extensions that are needed such that the proposed Stable Sparse PCA objective, optimizes concurrently both for the stability of the instance and feature clusters. We will furtheron refer to this type of concurrent stability optimization as "two-way" stability. As we illustrate in the following lemma, two-way stability can be achieved by employing double-centering, a popular data processing technique. Double centering essentially centers both rows and the columns of the data matrix such that they have zero mean. Based on double-centering the stability between the instance-clusters and feature clusters becomes equivalent. This effect is demonstrated in the following lemma whose proof can be found in the appendix.

Lemma 1. *Let X be our input $m \times n$ feature-instance data matrix. If X is double-centered, then the stability of spectral k-means applied on the instances is equivalent to the stability of spectral k-means applied on the features.*

Based on this observation we can extend the Stable Sparse PCA objective such that it optimizes for two-way stability. In order to achieve this goal we will define the double centered covariance matrix (omitting again the $1/n$ factor) as:

$$Cov = C_m^u X_{fc} X_{fc}^T C_m^u \tag{4}$$

The notation is the same as in the previous section, with the addition of C_m^u that is a matrix that performs instance-centering after feature selection, i.e. it is defined as $C_m^u = \mathbf{diag}(u)(I - \frac{1}{\mathbf{card}(u)} e_m e_m^T)\mathbf{diag}(u)$. It can be observed that if we consider the multiplication $C_m^u X$, the instances (columns) of matrix X are centered *after* the removal of features that correspond to $u(i) = 0$. The two-way stability optimizing objective is now defined simply by replacing the new Cov matrix in the optimization problem 2. Having defined the two-way stable Sparse PCA objective, we will move on in the next section for defining the appropriate efficient optimization framework for performing feature selection.

4 Optimization Framework

4.1 Useful Bounds for Optimizing Stability

Sparse PCA problems are known to be computationally hard and several approximation schemes have been developed for tackling them. In the context of this work we adopt the general approach of performing a greedy forward search that optimizes a lower bound of the stability maximizing objective. This general approach has been also adopted by other Sparse PCA algorithms (such as [3]). The derived bound is summarized in Theorem 1. In the theorem statement we use the same notation as in Section 3.1: **card** denotes the cardinality of a set, $x_{fc}(i)$ denotes the centered representation of a feature

and C_m^u is a matrix that performs instance-centering after feature selection, i.e. it is defined as $C_m^u = \mathbf{diag}(u)(I - \frac{1}{\mathbf{card}(u)} e_m e_m^T)\mathbf{diag}(u)$. The bound is derived for the more complex two-way stable objective. Based on the proof, a simpler bound for the one-way stability case can also be obtained.

Theorem 1. *Let I be a set of features and m a feature such that m does not belong to set I. Moreover, let v denote the dominant eigenvector of matrix $X_{fc}^T C_m^u X_{fc}$ as computed using features in set I. Then, the following lower bound can be derived:*

$$Obj(I \cup \{m\}) \geq Obj(I) + B$$

where

$$\begin{aligned} B = (1 - \tfrac{1}{\boldsymbol{card}(I)+1})[(v^T x_{fc}(m)^2 & - \tfrac{1}{n} x_{fc}(m)^T x_{fc}(m)] \\ - \tfrac{2}{\boldsymbol{card}(I)+1}[(\textstyle\sum_{i\in I} v^T x_{fc}(i)) v^T x_{fc}(m) & - \tfrac{1}{n}(\textstyle\sum_{i\in I} x_{fc}(i))^T x_{fc}(m)] \\ + \tfrac{1}{\boldsymbol{card}(I)(\boldsymbol{card}(I)+1)}[(v^T \textstyle\sum_{i\in I} x_{fc}(i))^2 & - \tfrac{1}{n}(\textstyle\sum_{i\in I} x_{fc}(i))^T (\textstyle\sum_{i\in I} x_{fc}(i))] \end{aligned}$$

It can be observed that the computational cost of this bound is dominated by the cost of computing the dominant eigenvector v of matrix $X_{fc}^T C_m^u X_{fc}$. The suitability of this bound for selecting the feature subset that maximizes for two-way stability is illustrated in the experiments section.

4.2 Greedy Solutions

In order to design efficient approximation schemes for the Stable Sparse PCA objective, we turn to greedy approaches. The proposed greedy algorithm is essentially an adaptation of the greedy strategies proposed in [3] that takes into account for the two distinct elements of our framework (double-centering and two-way stability). Our greedy algorithm also takes advantage of the lower bound derived in Theorem 1 and performs the greedy search *without explicitly computing the objective function* for each candidate feature.

The complexity of Algorithm 1 is $O(np^3 + n^2p^2 + nm^2)$ where p is the number of selected features and n the number of instances. This is because, at each step l (when selecting the l^{th} feature) in order to compute the bound we must double center the data matrix $O(nl^2 + n^2 l)$ (complexity of double centering an $n \times l$ matrix) and then compute the maximum eigenvalue of a matrix of size $n \times n$ which is $O(n^2)$ only once per greedy step. The candidate feature is selected based on the maximum angle between certain vector-pairs of sizes $n \times 1$ that induce a computational cost of $O(n(m-l))$. Since, double centering and the maximum eigenvalue is computed only once per greedy step of the algorithm, the total complexity will be $O(np^3 + n^2p^2 + nm^2)$.

It should be noted that because of double centering, the proposed algorithm is not able to select the initial feature (all features would appear to have quality equal to zero), thus the greedy algorithm is initialized with the feature that maximizes O_1 component of the objective (as defined within the proof of Theorem 1). It can be easily observed that this will essentially be the feature that has maximum variance. The algorithm terminates when the desired number of features p is selected.

Algorithm 1. (X,p)

1: Initialize with index I_{k_0} where $i_0 = argmax_{j \in I} O_1\{j\}$.
2: **repeat**
3: Compute $i_k = argmax_{i \in I_k^c} B(i,I_k)$.($B(i,I_k)$ is the lower bound of theorem 1)
4: Set $I_{k+1} = I_k \cup \{i_k\}$.
5: **until card**$(I_{k+1}) = p$.

4.3 Efficient Deflation for Multiple Clusters

In order to extend our framework for multiple clusters ($k > 2$), we consider the use of deflation. Although deflation is a rather straight forward approach for extracting multiple eigenvectors in the full feature case, it presents certain challenges in the context of sparse methods. These challenges are analytically illustrated in [11] where several deflation methods and their properties were thoroughly analyzed. Based on [11], one could simply employ one of the proposed methods, such as the Schur complement deflation, for computing the sequential sparse eigenvectors of the Covariance matrix. One issue with employing an "of-the-shelve" approach is that we would need to compute the Cholesky decomposition of the covariance matrix, in order to derive the new centered feature representations that are consequently employed in the bound computations (i.e. the $x_{fc}(i)$ in Theorem 1). This would affect the computational cost of the proposed method as it would include a $O(m^3)$ term for the Cholesky decomposition. In order to avoid this computational cost we propose a deflation process that is directly applied on the centered feature matrix X_{fc} using the dominant eigenvector of matrix $X_{fc}^T C_m^u X_{fc}$. As we will demonstrate, the proposed deflation is essentially equivalent to Schur complement deflation.

$$X_{fc}^{(t)} = X_{fc}^{(t-1)}(I - v_t v_t^T) \tag{5}$$

Starting from $t = 0$, the original input matrix of centered features, $X_{cf}^{(0)}$ is used for computing $Cov^{(0)}$ and for deriving the subset of features (as encoded in $u(t=0) \in \{0,1\}^m$) that optimizes the Stable Sparse PCA objective. Based on the selected features, the dominant eigenvector v_1 of $(X_{fc}^{(0)})^T C_m^{u(0)} X_{fc}^{(0)}$ is derived. Consecutively, we can employ the deflation formula for computing all the necessary eigenvectors. An interesting property of the deflation process is that at each sequential step, a different feature subset may be derived, thus giving the flexibility to the feature selection algorithm to select different feature subsets for separating between different clusters.

In order to illustrate the appropriateness of the afore proposed deflation method, we demonstrate that it is equivalent to a Schur complement deflation. The proof of this theorem can be found in the appendix.

Theorem 2. *The deflation procedure defined in equation 5 is equivalent to a Schur complement deflation on the feature covariance matrix.*

5 Related Work

The proposed framework is conceptually related to the work [13] that attributes the success of feature selection methods to the reduction of the data model variance. Under this

approach, features should not be selected simply by assessing their relevance to the target class but by considering the contribution that the features have to the bias-variance tradeoff of the learned model. That is, weakly relevant features that contribute much to the variance of the model should be excluded, while borderline-relevant features with low variance contribution can present good candidates for inclusion. In our study we adopt this principle and derive a criterion that selects features based on their contribution to cluster separation, weighted against their variance. Interestingly the cluster-separation variance tradeoff is derived through a stability maximizing objective.

In the relevant data mining literature, the term "stability of feature selection" is employed in a different manner and commonly refers to the robustness of the feature process itself, i.e. the ability of a feature selection algorithm to select the same feature with respect to noise, or data sample variations. The intuitiveness of this requirement has resulted in several works that study the stability of feature selection algorithms [9,7,15,10,21]. Our work is substantially different from these approaches, since it focuses on the effect of feature selection to the stability of the clustering output and not the feature selection process itself. Albeit this important differentiation, the "two-way" stability optimization framework can be employed in conjunction with the works [21,10]. This is because these methods employ the clustering structure of the features and perform feature selection at the clustering level (i.e. they select the relevant/stable feature clusters). In this context our approach can be employed as a preprocessing step that stabilizes the feature clusters thus enhancing the robustness of these methods.

The idea of selecting the features that optimize for an eigengap of a certain input matrix was also put forward in [20]. In this work the authors propose a continuous feature weighting scheme that achieves sparsity in an indirect manner. As opposed to [20] we conduct a detailed analysis on the properties of a stability maximizing objective in the context of k-means clustering. Moreover, we explicitly formulate our algorithm as a discrete feature selection and propose a novel Sparse PCA approach for selecting the appropriate features.

In [12] the authors optimize for stability by removing the features that contribute maximally to the variance of the derived model. This approach does not take into account the relevance of each feature and thus high quality features may be removed. In contrast, our approach explicitly takes into account the cluster separation quality of each feature that is weighted against the feature-variance.

In order to empirically validate our approach we compare our approach to the popular Laplacian score [8] and the recently proposed MCFS algorithm [1]. Although these algorithms are not conceptually relevant to the proposed framework, they provide a good basis for demonstrating that the proposed feature selection framework can achieve comparable performance with state-of-the-art algorithms.

6 Experiments

In the context of the Cancer research application, we have experimented with four publicly available microarray datasets that were obtained from `http://algorithmics.molgen.mpg.de/Static/Supplements/CompCancer/datasets.htm`. The employed datasets are summarized in the following table:

Name	Description	#Instances	#Features	#Classes
Chen-2002	Liver Cancer	179	85	2
Golub-1999-v2	Leukemia	72	1877	3
Pomeroy-2002-v2	Central Nervous System Tumors	42	1379	5
Ramaswamy-2001	Multiple Cancer Types	190	1363	14

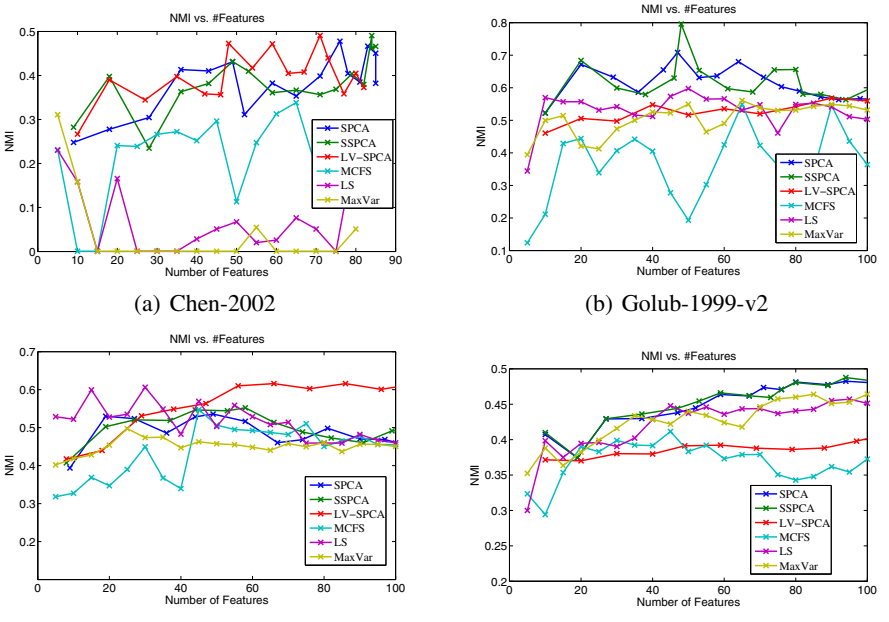
We compare our feature selection framework against *Laplacian Score* [8], the recently proposed *MCFS* algorithm [1] and the simple heuristic of selecting the features that have maximal variance. In order to conduct the comparison, we employ the selected feature subsets in the context of a k-means clustering algorithm and compute the achieved cluster quality using Normalized Mutual Information *(NMI)*.

It should be noted that our method differs substantially from the *Laplacian Score* and *MCFS*. The main difference is that our method is "faithful" to the k-means objective, while the methods we compare against construct a Laplacian matrix for measuring the relevance of the features. It is evident that the construction of the Laplacian matrix can substantially influence the results (i.e. by considering different similarity functions, Gaussian vs. simple inner-product or different types of graphs, k-nn Graphs vs. Full Graphs). Thus, a major factor that determines which method is more suitable for different application contexts depends on the ability to construct/tune an appropriate Laplacian matrix and also on the properties of the underlying clustering structure of the data.

For constructing the Laplacian matrix for the *MCFS* and *Laplacian Score* we employ the cosine similarity for computing the instance-similarity matrix W and then construct a k-nearest neighbor graph with $k = 5$. These settings are similar to the ones recommended in [1]. Moreover, the number of Laplacian eigenvectors that were employed within *MCFS* was set to be equal to the number of clusters.

Our method constructs the continuous cluster indicators using 2 sparse eigenvectors in all datasets (instead of #Cluster-1). In three out of four dataset this is different than the (#Cluster-1) that is recommended by the "pure" Spectral k-means approach. In two cases (Pomeroy and Ramaswamy datasets) we have employed only two sparse eigenvectors in order to obtain comparable results for a small number of features. This is because the proposed framework can select different features for each eigenvector, thus even when selecting a small number of features for each sparse eigenvector, the total number of features can be much larger. In the Chen-2002 dataset we employed 2 sparse eigenvectors for performance reasons, since we observed that using only 1 eigenvector did not suffice to obtain a good clustering performance. We should finally note that the standard Lloyd's k-means algorithms was used for discretizing the continuous results and also that we have employed the "two-way" stable version of our objective.

In the experiments we have also explored the possibilities of introducing different tradeoffs between cluster-separation and feature variance. More precisely we have experimented with three objectives: The standard cluster-separation vs. variance tradeoff that is derived by maximizing the average eigengap between the largest and the consecutive eigenvalues. $\lambda_1(Cov) - \frac{1}{n}\mathbf{Trace}(Cov)$ (denoted in the Figures as *SSPCA*), a "pure" Sparse PCA approach that maximizes solely $\lambda_1(Cov)$ (denoted in the Figures as *SPCA*), and a "low variance" approach that penalizes heavily variance using $\lambda_1(Cov) - \gamma\mathbf{Trace}(Cov)$, where $\gamma = \frac{\lambda_1(Cov_{full})}{\mathbf{Trace}(Cov_{full})}$ i.e. it is equal to the ratio of the



(a) Chen-2002 (b) Golub-1999-v2

(c) Pomeroy-2002-v2 (d) Ramaswamy-2001

Fig. 1. Comparative Study of Clustering Quality

maximum eigenvalue to the Trace of the original full-feature Covariance matrix, before feature selection (denoted in the Figures as *LV-SPCA*).

In Figure 1 we can observe that the clustering quality is competitive and often superior against the relevant feature selection methods. More precisely, in all Figures at least one of the three Sparse PCA methods is superior with the exception of Figure 1(c), where the Laplacian Score is better for small feature sizes. Moreover, we can observe a mixed behavior with respect to the appropriate level of variance penalization, with *LV-SPCA* demonstrating a very good performance in two out of four experiments.

Apart from the indirect evaluation of the accuracy of feature selection (through clustering quality) we also investigate whether the selected features can provide insights into the underlying problem under study. For this purpose we focus on the Golub-1999-v2 dataset [6] that is related to Acute lymphoblastic leukemia (ALL), which is the most common pediatric cancer, accounting for 30% of all pediatric malignancies.

Golub's dataset [6] consists of bone marrow sample from acute leukemia patients, involving myeloid leukemia (AML) and two sub-types of acute lymphoblastic leukemia (ALL), B-cell and T-cell ALL. The analyzed Golub-1999-v2 dataset consists of 38 ALL-B, 9 ALL-T and 25 AML samples and 1877 genes.

We compared the top 5 features selected in the first and second eigenvector generated using two different versions of the proposed algorithm *SPCA* (where solely $\lambda_1(Cov)$ is optimized with no variance correction) and *LV-SPCA* that employs a strong variance correction threshold. Table 1 shows the list of genes. As perhaps one could expect, the variance of the genes only selected using *LV-SPCA* is in general smaller than the one of

those selected using *SPCA*. In order to investigate the relevance of the selected genes for the disease under study, we performed a literature research on the three genes uniquely identified by the proposed method with direct optimization of stability.

Genes Uniquely Selected by LV-SPCA. Gene number 458 has ID M21005_at and corresponds to the S100 calcium binding protein A8 (calgranulin A). A very recent experimental investigation has been performed in [14] which suggests that the expression of S100A8 in leukemic cells is a predictor of low survival. This gene was not found among top 100 of *MCFS*, and was ranked by *Laplacian Score* as 66th. Furthermore, this gene was also not predicted as relevant by other gene ranking methods (see *http://genomics10.bu.edu/yangsu/rankgene/compare-ALL-AML-all-top100.html#ranks_table*).

Gene number 1614 has ID Y00787_s_at and corresponds to the Interleukin-8 precursor. In [16] it has been suggested that Interleukin-8 upregulation may play a role in the pathogenesis of T-cell acute lymphoblastic leukemia.

Gene number 1613 has ID M28130_rna1_s_at and corresponds to the Interleukin 8 (IL8) gene. It has been suggested that IL-8 may function as a significant regulatory factor within the tumor microenvironment. Recently, IL-8 signaling has been implicated in regulating the transcriptional activity of the androgen receptor, underpinning the transition to an androgen-independent proliferation of prostate cancer cells. In addition, stress and drug-induced IL-8 signaling has been shown to confer chemotherapeutic resistance in cancer cells. Therefore, inhibiting the effects of IL-8 signaling may be a significant therapeutic intervention in targeting the tumor microenvironment [19]. Indeed, Interleukin 8 (IL-8) is currently being applied in various subspecialties of medicine either for the purpose of rapid diagnosis or as a predictor of prognosis: in [17] an overview of current evidence is provided suggesting that Interleukin 8 (IL-8) may serve as a useful biomarker.

Genes Uniquely Selected by SPCA. Gene number 607 has ID M91036_rna1_at and corresponds to the G-gamma globin gene. Gene number 1798 has ID U01317_cds4_at and corresponds to the Delta-globin gene. We did not find strong evidence of a relation of these two genes with the leukemia pathogenesis and pharmacology.

Gene number 493 has ID U10685_at and corresponds to the MAGE A10 gene. The mammalian members of the MAGE (melanoma-associated antigen) gene family were originally described as completely silent in normal adult tissues, with the exception of male germ cells and, for some of them, placenta. By contrast, these genes were expressed in various kinds of tumors. However, other members of the family were recently

Table 1. Index of top 5 genes selected by the method in the first and second eigenvector for *SPCA* and *LV-SPCA*. The number between brackets indicates the position of the gene in the list of gene sorted in decreasing with respect to the variance.

Method	Eigenvector	Feat1	Feat2	Feat3	Feat4	Feat5
SPCA	i=1	1623 (1)	1194 (5)	**493** (2)	1106 (17)	672 (33)
SPCA	i=2	**607** (8)	1734 (9)	435 (15)	**1798** (16)	1756 (27)
LV-SPCA	i=1	1623 (1)	1194 (5)	**458** (18)	672 (33)	1106 (17)
LV-SPCA	i=2	**1614** (4)	1734 (9)	**1613** (23)	435 (15)	1756 (27)

found to be expressed in normal cells, indicating that the family is larger and more disparate than initially expected [2].

The above observations indicate the effectiveness of optimizing stability for unsupervised feature selection with respect to the detection of features strongly related to the pathogenesis and pharmacology of the disease under study.

7 Conclusions and Further Work

In conclusion, we have proposed a novel feature selection framework that maximizes the stability of Spectral k-means. The semantics of the proposed framework are analyzed in detail and it is demonstrated that stability maximization naturally leads to a cluster-separation vs. variance tradeoff. As a matter of further work, we aim in extending our framework to Kernel k-means and also to Spectral Clustering algorithms that employ the Laplacian matrix.

Appendix

Proof (Proof of Lemma 1). Based on [5], the continuous solution for the instance clusters is derived by the $k-1$ dominant eigenvectors of matrix $X_{fc}^T X_{fc}$, where X_{fc} is a feature-instance matrix with the rows (features) being centered. Since X is double-centered the sum of rows and columns of X will be equal to 0, i.e. $\sum_i X_{ij} = \sum_j X_{ij} = 0$. Thus, the continuous solution of Spectral k-means (for instance clustering) will be derived by the $k-1$ dominant eigenvectors of matrix $X^T X$.

Analogously, the continuous solution for the feature clusters is derived by the $k-1$ dominant eigenvectors of matrix $X_{ic} X_{ic}^T$, where X_{ic} is a feature-instance matrix with the columns (instances) being centered. Since X is double-centered, we will have that the instance-centered matrix X_{ic} will be equal to X, i.e. $X_{ic} = X$. Thus, the continuous cluster solution will be derived by the dominant eigenvectors of matrix XX^T.

Using basic linear algebra one can easily derive that the matrices XX^T and $X^T X$ have exactly the same eigenvalues.

Thus $\lambda_{k-1}(XX^T) - \lambda_k(XX^T) = \lambda_{k-1}(X^T X) - \lambda_k(X^T X)$, and the stability of the relevant eigenspaces will be equivalent.

Proof (Proof of Theorem 1). We will start by decomposing the components $\lambda_1(Cov)$ and $\mathbf{Trace}(Cov)$.

For $\lambda_1(Cov)$ we have:

$$\begin{aligned}
\lambda_1(Cov) &= \lambda_1(C_m^u X_{fc} X_{fc}^T C_m^u) \\
&= \lambda_1(X_{fc}^T C_m^u X_{fc}) \\
&= \lambda_1(X_{fc}^T \mathbf{diag}(u)(I - \tfrac{1}{\mathbf{card}(u)} e_m e_m^T)\mathbf{diag}(u) X_{fc}) \\
&= \max_{||v||=1} v^T (X_{fc}^T \mathbf{diag}(u)(I - \tfrac{1}{\mathbf{card}(u)} e_m e_m^T)\mathbf{diag}(u) X_{fc}) v \\
&= \max_{||v||=1} v^T (X_{fc}^T \mathbf{diag}(u) X_{fc}) v \\
&\qquad - \tfrac{1}{\mathbf{card}(u)} v^T (X_{fc}^T \mathbf{diag}(u) e_m e_m^T \mathbf{diag}(u) X_{fc}) v) \\
&= \max_{||v||=1} \sum_{i=1}^{m} u_i (v^T x_{fc}(i))^2 - \tfrac{1}{\mathbf{card}(u)} (v^T \sum_{i=1}^{m} u_i x_{fc}(i))^2
\end{aligned}$$

In the above derivations we have used the following, easily verifiable facts: $\lambda_1(AA^T) = \lambda_1(A^TA)$, $C_m^u = (C_m^u)^T$, $C_m^u = (C_m^u)^2$ and $\mathbf{diag}(u) = (\mathbf{diag}(u))^2$.

For $\mathbf{Tr}(Cov)$ we have:

$$\begin{aligned}\mathbf{Tr}(Cov) &= \mathbf{Tr}(C_m^u X_{fc} X_{fc}^T C_m^u)\\ &= \mathbf{Tr}(X_{fc}^T C_m^u X_{fc})\\ &= \mathbf{Tr}(X_{fc}\mathbf{diag}(u)(I - \tfrac{1}{\mathbf{card}(u)} e_m e_m^T)\mathbf{diag}(u) X_{fc})\\ &= \mathbf{Tr}(X_{fc}^T \mathbf{diag}(u) X_{fc})\\ &\quad - \tfrac{1}{\mathbf{card}(u)}\mathbf{Tr}(X_{fc}^T \mathbf{diag}(u) e_m e_m^T \mathbf{diag}(u) X_{fc})\end{aligned}$$

In these derivations we have used the following properties of the matrix Trace: $\mathbf{Tr}(AA^T) = \mathbf{Tr}(A^TA)$, $\mathbf{Tr}(A+B) = \mathbf{Tr}(A) + \mathbf{Tr}(B)$ and $\mathbf{Tr}(\beta A) = \beta \mathbf{Tr}(A)$.

Now for $\mathbf{Tr}(X_{fc}^T \mathbf{diag}(u) X_{fc})$ we have:

$$\begin{aligned}\mathbf{Tr}(X_{fc}^T \mathbf{diag}(u) X_{fc}) &= \mathbf{Tr}(\mathbf{diag}(u) X_{fc} X_{fc}^T \mathbf{diag}(u))\\ &= \sum_{i=1}^{m} u_i x_{fc}(i)^T x_{fc}(i)\end{aligned}$$

Finally, for $\frac{1}{\mathbf{card}(u)}\mathbf{Tr}(X_{fc}^T \mathbf{diag}(u) e_m e_m^T \mathbf{diag}(u) X_{fc})$ we have:

$$\begin{aligned}&\tfrac{1}{\mathbf{card}(u)}\mathbf{Tr}(X_{fc}^T \mathbf{diag}(u) e_m e_m^T \mathbf{diag}(u) X_{fc}) =\\ &\tfrac{1}{\mathbf{card}(u)}\mathbf{Tr}(e_m^T \mathbf{diag}(u) X_{fc} X_{fc}^T \mathbf{diag}(u) e_m) =\\ &\tfrac{1}{\mathbf{card}(u)}(\sum_{i=1}^{n} u_i x_{fc}(i))^T (\sum_{i=1}^{m} u_i x_{fc}(i))\end{aligned}$$

Based on the afore derivations we can write the objective function as:

$$\max_{||v||=1} \max_{u \in \{0,1\}^n} [O_1 - \frac{1}{\mathbf{card}(u)} O_2] \tag{6}$$

Where

$$O_1 = \sum_{i=1}^{m} u_i [(v^T x_{fc}(i))^2 - \frac{1}{n} x_{fc}(i)^T x_{fc}(i)]$$

$$O_2 = (v^T \sum_{i=1}^{m} u_i x_{fc}(i))^2 - \frac{1}{n}(\sum_{i=1}^{m} u_i x_{fc}(i))^T (\sum_{i=1}^{m} u_i x_{fc}(i))$$

Based on the derivation of the objective function using O_1 and O_2 and also the fact that for all v such that $||v|| = 1$ it holds that $\lambda_1(A) \geq x^T A x$, the lower bound can be derived.

Proof (Proof of Theorem 2). Recall that in Schur complement deflation, the deflation step is performed as follows:

$$A_t = A_{t-1} - \frac{A_{t-1} x_t x_t^T A_{t-1}}{x_t^T A_{t-1} x_t}$$

Now if we consider that $A_t = X_{fc}^{(t)} (X_{fc}^{(t)})^T$ and also that x_t is the dominant eigenvector of matrix $Cov^{(t-1)}$, we can write:

$$A_t = A_{t-1} - \frac{A_{t-1}x_t x_t^T A_{t-1}}{x_t^T A_{t-1} x_t} \Rightarrow$$

$$X_{fc}^{(t)}(X_{fc}^{(t)})^T = X_{fc}^{(t-1)}(X_{fc}^{(t-1)})^T - \frac{X_{fc}^{(t-1)}(X_{fc}^{(t-1)})^T x_t x_t^T X_{fc}^{(t-1)}(X_{fc}^{(t-1)})^T}{x_t^T X_{fc}^{(t-1)}(X_{fc}^{(t-1)})^T x_t} \tag{7}$$

Recall that x_t is the dominant eigenvector of $Cov^{(t-1)}$ that can be written as $Cov^{(t-1)} = C_m^{u(t-1)} X_{fc}^{(t-1)}(X_{fc}^{(t-1)})^T C_m^{u(t-1)}$ (i.e. it is based on the selected feature subset $u(t-1)$).

Since $Cov^{(t-1)}$ is a double-centered matrix (its rows and columns are centered through the multiplication with $C_m^{u(t-1)}$), its dominant eigenvector will also be centered, i.e. $(C_m^{u(t-1)})x_t = x_t$. Based on this property, we can write:

$$\begin{aligned} x_t^T X_{fc}^{(t-1)}(X_{fc}^{(t-1)})^T x_t &= x_t^T C_m^{u(t-1)} X_{fc}^{(t-1)}(X_{fc}^{(t-1)})^T C_m^{u(t-1)} x_t \\ &= x_t^T Cov^{(t-1)} x_t = \lambda_{max}^{(t-1)} \end{aligned} \tag{8}$$

Now, $(X_{fc}^{(t-1)})^T x_t$ can be written as:

$$(X_{fc}^{(t-1)})^T x_t = (X_{fc}^{(t-1)})^T C_m^{u(t-1)} x_t = \sqrt{\lambda_{max}^{(t-1)}} v_t, \tag{9}$$

where v_t is the dominant eigenvector of $(X_{fc}^{(t-1)})^T C_m^{u(t-1)} X_{fc}^{(t-1)}$ and $\lambda_{max}^{(t-1)}$ is the dominant eigenvector of $Cov^{(t-1)}$.

Based on equations 7,8,9 we can write

$A_t = A_{t-1} - \frac{A_{t-1}x_t x_t^T A_{t-1}}{x_t^T A_{t-1} x_t} \Rightarrow$
$X_{fc}^{(t)}(X_{fc}^{(t)})^T = X_{fc}^{(t-1)}(X_{fc}^{(t-1)})^T - X_{fc}^{(t-1)} v_t v_t^T (X_{fc}^{(t-1)})^T \Rightarrow$
$X_{fc}^{(t)} = X_{fc}^{(t-1)}(I - v_t v_t^T)$

References

1. Cai, D., Zhang, C., He, X.: Unsupervised feature selection for multi-cluster data. In: ACM SIGKDD (2010)
2. Chomez, P., Backer, O.D., Bertrand, M., Plaen, E.D., Boon, T., Lucas, S.: An overview of the mage gene family with the identification of all human members of the family. Cancer Research 15, 6 (2001)
3. d'Aspremont, A., Bach, F.R., Ghaoui, L.E.: Full regularization path for sparse principal component analysis. In: ICML (2007)
4. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: ACM SIGKDD (2001)
5. Ding, C.H.Q., He, X.: K-means clustering via principal component analysis. In: ICML (2004)
6. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439), 531–537 (1999)
7. Han, Y., Yu, L.: A variance reduction framework for stable feature selection. In: IEEE ICDM (2010)

8. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: NIPS (2005)
9. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. 12(1), 95–116 (2007)
10. Loscalzo, S., Yu, L., Ding, C.H.Q.: Consensus group stable feature selection. In: ACM SIGKDD (2009)
11. Mackey, L.: Deflation methods for sparse pca. In: NIPS (2008)
12. Mavroeidis, D., Vazirgiannis, M.: Stability based sparse LSI/PCA: Incorporating feature selection in LSI and PCA. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 226–237. Springer, Heidelberg (2007)
13. Munson, M.A., Caruana, R.: On feature selection, bias-variance, and bagging. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009. LNCS, vol. 5782, pp. 144–159. Springer, Heidelberg (2009)
14. Nicolas, E., Ramus, C., Berthier, S., Arlotto, M., Bouamrani, A., Lefebvre, C., Morel, F., Garin, J., Ifrah, N., Berger, F., Cahn, J.Y., Mossuz, P.: Expression of s100a8 in leukemic cells predicts poor survival in de novo aml patients. Leukemia 25, 57–65 (2011)
15. Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 313–325. Springer, Heidelberg (2008)
16. Scupoli, M., Donadelli, M., Cioffi, F., Rossi, M., Perbellini, O., Malpeli, G., Corbioli, S., Vinante, F., Krampera, M., Palmieri, M., Scarpa, A., Ariola, C., Foa, R., Pizzolo, G.: Bone marrow stromal cells and the upregulation of interleukin-8 production in human t-cell acute lymphoblastic leukemia through the cxcl12/cxcr4 axis and the nf-kappab and jnk/ap-1 pathways. Haematologica 93(4), 524–532 (2008)
17. Shahzad, A., Knapp, M., Lang, I., Kohler, G.: Interleukin 8 (il-8) - a universal biomarker? International Archives of Medicine 3(11) (2010)
18. Stewart, G.W., Sun, J.G.: Matrix Perturbation Theory (Computer Science and Scientific Computing). Academic Press, London (1990)
19. Waugh, D., Wilson, C.: The interleukin-8 pathway in cancer. Clinical Cancer Research (2008)
20. Wolf, L., Shashua, A.: Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach. J. Mach. Learn. Res. (2005)
21. Yu, L., Ding, C.H.Q., Loscalzo, S.: Stable feature selection via dense feature groups. In: ACM SIGKDD (2008)