Randomized Dimensionality Reduction for k-Means Clustering

Christos Boutsidis, Anastasios Zouzias, Michael W. Mahoney, and Petros Drineas

***Abstract*— We study the topic of dimensionality reduction for k-means clustering. Dimensionality reduction encompasses the union of two approaches: 1) feature selection and 2) feature extraction. A feature selection-based algorithm for k-means clustering selects a small subset of the input features and then applies k-means clustering on the selected features. A feature extraction-based algorithm for k-means clustering constructs a small set of new artificial features and then applies k-means clustering on the constructed features. Despite the significance of k-means clustering as well as the wealth of heuristic methods addressing it, provably accurate feature selection methods for k-means clustering are not known. On the other hand, two provably accurate feature extraction methods for k-means clustering are known in the literature; one is based on random projections and the other is based on the singular value decomposition (SVD). This paper makes further progress toward a better understanding of dimensionality reduction for k-means clustering. Namely, we present the first provably accurate feature selection method for k-means clustering and, in addition, we present two feature extraction methods. The first feature extraction method is based on random projections and it improves upon the existing results in terms of time complexity and number of features needed to be extracted. The second feature extraction method is based on fast approximate SVD factorizations and it also improves upon the existing results in terms of time complexity. The proposed algorithms are randomized and provide constant-factor approximation guarantees with respect to the optimal k-means objective value.**

***Index Terms*— Clustering, dimensionality reduction, randomized algorithms.**

I. Introduction

CLUSTERING is ubiquitous in science and engineering with numerous application domains ranging from bioinformatics and medicine to the social sciences and the web [1]. Perhaps the most well-known clustering algorithm is the so-called "k-means" algorithm or Lloyd's method [2]. Lloyd's method is an iterative expectation-maximization type approach that attempts to address the following objective: given a set of Euclidean points and a positive integer k corresponding to the number of clusters, split the points into k clusters so that the total sum of the squared Euclidean distances of each point to its nearest cluster center is minimized. Due to this intuitive objective as well as its *effectiveness* [3], the Lloyd's method for k-means clustering has become enormously popular in applications [4].

In recent years, the high dimensionality of modern massive datasets has provided a considerable challenge to the design of efficient algorithmic solutions for k-means clustering. First, ultra-high dimensional data force existing algorithms for k-means clustering to be computationally inefficient, and second, the existence of many irrelevant features may not allow the identification of the relevant underlying structure in the data [5]. Practitioners have addressed these obstacles by introducing feature selection and feature extraction techniques. Feature selection selects a (small) subset of the actual features of the data, whereas feature extraction constructs a (small) set of artificial features based on the original features. Here, we consider a rigorous approach to feature selection and feature extraction for k-means clustering. Next, we describe the mathematical framework under which we will study such dimensionality reduction methods.

Consider m points $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m\} \subseteq \mathbb{R}^n$ and an integer k denoting the number of clusters. The objective of k-means is to find a k-partition of $\mathcal{P}$ such that points that are "close" to each other belong to the same cluster and points that are "far" from each other belong to different clusters. A k-partition of $\mathcal{P}$ is a collection $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k\}$ of k non-empty pairwise disjoint sets which covers $\mathcal{P}$. Let $s_j = |\mathcal{S}_j|$ be the size of $\mathcal{S}_j$ ($j = 1, 2, \ldots, k$). For each set S_j, let $\boldsymbol{\mu}_j \in \mathbb{R}^n$ be its centroid:

$$\boldsymbol{\mu}_j = \frac{\sum_{\mathbf{p}_i \in S_j} \mathbf{p}_i}{s_j}.$$

The k-means objective function is

$$\mathcal{F}(\mathcal{P}, \mathcal{S}) = \sum_{i=1}^{m} \|\mathbf{p}_i - \boldsymbol{\mu}(\mathbf{p}_i)\|_2^2,$$

where $\boldsymbol{\mu}(\mathbf{p}_i) \in \mathbb{R}^n$ is the centroid of the cluster to which $\mathbf{p}_i$ belongs. The objective of k-means clustering is to compute the optimal k-partition of the points in $\mathcal{P}$,

$$\mathcal{S}_{opt} = \arg\min_{\mathcal{S}} \mathcal{F}(\mathcal{P}, \mathcal{S}).$$

Now, the goal of dimensionality reduction for k-means clustering is to construct points

$$\hat{\mathcal{P}} = \{\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, \ldots, \hat{\mathbf{p}}_m\} \subseteq \mathbb{R}^r$$

Manuscript received October 15, 2013; revised June 16, 2014; accepted October 15, 2014. Date of publication November 26, 2014; date of current version January 16, 2015.

C. Boutsidis is with Yahoo Labs. New York, NY 10036 USA (e-mail: boutsidis@yahoo-inc.com).

A. Zouzias is with the Department of Computer Science, University of Toronto, Toronto, ON M5S 2J7, Canada, and also with the IBM Research Laboratory, Zurich 8803, Switzerland (e-mail: azo@zurich.ibm.com).

M. W. Mahoney is with the Department of Statistics, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: mmahoney@stat.berkeley.edu).

P. Drineas is with the Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: drinep@cs.rpi.edu).

Communicated by N. Cesa-Bianchi, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2014.2375327

(for some parameter $r \ll n$) so that $\hat{\mathcal{P}}$ approximates the clustering structure of $\mathcal{P}$. Dimensionality reduction via feature selection constructs the $\hat{\mathbf{p}}_i$'s by selecting actual features of the corresponding $\mathbf{p}_i$'s, whereas dimensionality reduction via feature extraction constructs new artificial features based on the original features. More formally, assume that the optimum k-means partition of the points in $\hat{\mathcal{P}}$ has been computed

$$\hat{\mathcal{S}}_{opt} = \arg\min_{\mathcal{S}} \mathcal{F}(\hat{\mathcal{P}}, \mathcal{S}).$$

A dimensionality reduction algorithm for k-means clustering constructs a new set $\hat{\mathcal{P}}$ such that

$$\mathcal{F}(\mathcal{P}, \hat{\mathcal{S}}_{opt}) \leq \gamma \cdot \mathcal{F}(\mathcal{P}, \mathcal{S}_{opt})$$

where $\gamma > 0$ is the approximation ratio of $\hat{\mathcal{S}}_{opt}$. In other words, we require that computing an optimal partition $\hat{\mathcal{S}}_{opt}$ on the projected low-dimensional data and plugging it back to cluster the high dimensional data, will imply a γ factor approximation to the optimal clustering. Notice that we measure approximability by evaluating the k-means objective function, which is a well studied approach in the literature [3], [6]–[11]. Comparing the structure of the actual clusterings $\hat{\mathcal{S}}_{opt}$ to $\mathcal{S}_{opt}$ would be much more interesting but our techniques do not seem to be helpful towards this direction. However, from an empirical point of view (see Section VII), we do compare $\hat{\mathcal{S}}_{opt}$ directly to $\mathcal{S}_{opt}$ observing favorable results.

A. Prior Work

Despite the significance of dimensionality reduction in the context of clustering, as well as the wealth of heuristic methods addressing it [14], to the best of our knowledge there are no provably accurate feature selection methods for k-means clustering known. On the other hand, two provably accurate feature extraction methods are known in the literature that we describe next.

First, a result by [15] indicates that one can construct $r = O(\log(m)/\varepsilon^2)$ artificial features with Random Projections and, with high probability, obtain a $(1+\varepsilon)$-approximate clustering. The algorithm implied by [15], which is a random-projection-type algorithm, is as follows: let $\mathbf{A} \in \mathbb{R}^{m \times n}$ contain the points $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m\} \subseteq \mathbb{R}^n$ as its rows; then, multiply $\mathbf{A}$ from the right with a random projection matrix $\mathbf{R} \in \mathbb{R}^{n \times r}$ to construct $\mathbf{C} = \mathbf{AR} \in \mathbb{R}^{m \times r}$ containing the points $\hat{\mathcal{P}} = \{\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, \ldots, \hat{\mathbf{p}}_m\} \subseteq \mathbb{R}^r$ as its rows (see Section III-B for a definition of a random projection matrix). The proof of this result is immediate mainly due to the Johnson-Lindenstrauss lemma [15]. Ref. [15] proved that all the pairwise Euclidean distances of the points of $\mathcal{P}$ are preserved within a multiplicative factor $1 \pm \varepsilon$. So, any value of the k-means objective function, which depends only on pairwise distances of the points from the corresponding center point, is preserved within a factor $1 \pm \varepsilon$ in the reduced space.

Second, [12] argues that one can construct $r = k$ artificial features using the SVD, in $O(mn \min\{m, n\})$ time, to obtain a 2-approximation on the clustering quality. The algorithm of [12] is as follows: given $\mathbf{A} \in \mathbb{R}^{m \times n}$ containing the points of $\mathcal{P}$ and k, construct $\mathbf{C} = \mathbf{AV}_k \in \mathbb{R}^{m \times k}$. Here, $\mathbf{V}_k \in \mathbb{R}^{n \times k}$ contains the top k right singular vectors of $\mathbf{A}$. The proof of this result will be (briefly) discussed in Sections II-B and VI.

Finally, an extension of the latter SVD-type result (see [13, Corollary 4.5]) argues that $O(k/\varepsilon^2)$ dimensions (singular vectors) suffice for a relative-error approximation.

Algorithm 1 Randomized Feature Selection for k-Means Clustering

Input: Dataset $\mathbf{A} \in \mathbb{R}^{m \times n}$, number of clusters k, and $0 < \varepsilon < 1/3$.
Output: $\mathbf{C} \in \mathbb{R}^{m \times r}$ with $r = O(k \log(k)/\varepsilon^2)$ rescaled features.

1: Let $\mathbf{Z} = \text{FastFrobeniusSVD}(\mathbf{A}, k, \varepsilon)$; $\mathbf{Z} \in \mathbb{R}^{n \times k}$ (via Lemma 2).
2: Let $r = c_1 \cdot 4k \ln(200k)/\varepsilon^2$ (c_1 is a sufficiently large constant - see proof).
3: Let $[\mathbf{\Omega}, \mathbf{S}] = \text{RandomizedSampling}(\mathbf{Z}, r)$; $\mathbf{\Omega} \in \mathbb{R}^{n \times r}, \mathbf{S} \in \mathbb{R}^{r \times r}$ (via Lemma 3).
4: Return $\mathbf{C} = \mathbf{A\Omega S} \in \mathbb{R}^{m \times r}$ with r rescaled columns from $\mathbf{A}$.

B. Summary of Our Contributions

We present the first provably accurate feature selection algorithm for k-means (Algorithm 1). Namely, Theorem 1 presents an $O(mnk\varepsilon^{-1} + k \log(k)\varepsilon^{-2} \log(k \log(k)\varepsilon^{-1}))$ time randomized algorithm that, with constant probability, achieves a $(3+\varepsilon)$-error with $r = O(k \log(k)/\varepsilon^2)$ features. Given $\mathbf{A}$ and k, the algorithm of this theorem first computes $\mathbf{Z} \in \mathbb{R}^{n \times k}$, which approximates $\mathbf{V}_k \in \mathbb{R}^{n \times k}$ which contains the top k right singular vectors of $\mathbf{A}$.[1] Then, the selection of the features (columns of $\mathbf{A}$) is done with a standard randomized sampling approach with replacement with probabilities that are computed from the matrix $\mathbf{Z}$. The proof of Theorem 1 is a synthesis of ideas from [12] and [17], which study the paradigm of dimensionality reduction for k-means clustering and the paradigm of randomized sampling, respectively.

Moreover, we describe a random-projection-type feature extraction algorithm: Theorem 2 presents an $O(mn\lceil \varepsilon^{-2} k/\log(n) \rceil)$ time algorithm that, with constant probability, achieves a $(2+\varepsilon)$-error with $r = O(k/\varepsilon^2)$ artificial features. We improve the folklore result of the first row in Table I by means of showing that a smaller number of features are enough to obtain an approximate clustering. The algorithm of Theorem 2 is the same as with the one in the standard result for random projections that we outlined in the prior work section but uses only $r = O(k/\varepsilon^2)$ dimensions for the random projection matrix. Our proof relies on ideas from [12] and [18], which study the

[1] Ref. [16] presented an unsupervised feature selection algorithm by working with the matrix $\mathbf{V}_k$; in this work, we show that the same approximation bound can be achieved by working with a matrix that approximates $\mathbf{V}_k$ in the sense of low rank matrix approximations (see Lemma 2).

TABLE I

PROVABLY ACCURATE DIMENSIONALITY REDUCTION METHODS FOR k-MEANS CLUSTERING. RP STANDS FOR RANDOM PROJECTIONS, AND RS STANDS FOR RANDOM SAMPLING. THE THIRD COLUMN CORRESPONDS TO THE NUMBER OF SELECTED/EXTRACTED FEATURES; THE FOURTH COLUMN CORRESPONDS TO THE TIME COMPLEXITY OF EACH DIMENSIONALITY REDUCTION METHOD; THE FIFTH COLUMN CORRESPONDS TO THE APPROXIMATION RATIO OF EACH APPROACH

Reference	Description	Dimensions	Time = $O(x), x =$	Approximation Ratio
Folklore	RP	$O(\log(m)/\varepsilon^2)$	$mn\lceil\varepsilon^{-2}\log(m)/\log(n)\rceil$	$1+\varepsilon$
[12]	Exact SVD	k	$mn\min\{m,n\}$	2
[13]	Exact SVD	$O(k/\varepsilon^2)$	$mn\min\{m,n\}$	$1+\varepsilon$
Theorem 11	RS	$O(k\log(k)/\varepsilon^2)$	mnk/ε	$3+\varepsilon$
Theorem 12	RP	$O(k/\varepsilon^2)$	$mn\lceil\varepsilon^{-2}k/\log(n)\rceil$	$2+\varepsilon$
Theorem 13	Approx. SVD	k	mnk/ε	$2+\varepsilon$

paradigm of dimension reduction for k-means clustering and the paradigm of speeding up linear algebra problems, such as the low-rank matrix approximation problem, via random projections, respectively.

Finally, Theorem 3 describes a feature extraction algorithm that employs approximate SVD decompositions and constructs $r = k$ artificial features in $O(mnk/\varepsilon)$ time such that, with constant probability, the clustering error is at most a $2+\varepsilon$ multiplicative factor from the optimal. We improve the existing SVD dimensionality reduction method by showing that fast approximate SVD gives features that can do almost as well as the features from the exact SVD. Our algorithm and proof are similar to those in [12], but we show that one only needs to compute an approximate SVD of $\mathbf{A}$.

We summarize previous results as well as our results in Table I.

II. LINEAR ALGEBRAIC FORMULATION AND OUR APPROACH

A. Linear Algebraic Formulation of k-Means

From now on, we will switch to a linear algebraic formulation of the k-means clustering problem following the notation used in the introduction. Define the data matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ whose rows correspond to the data points,

$$\mathbf{A}^{\mathrm{T}} = [\mathbf{p}_1, \ldots, \mathbf{p}_m] \in \mathbb{R}^{n\times m}.$$

We represent a k-clustering $\mathcal{S}$ of $\mathbf{A}$ by its *cluster indicator matrix* $\mathbf{X} \in \mathbb{R}^{m\times k}$. Each column $j = 1, \ldots, k$ of $\mathbf{X}$ corresponds to a cluster. Each row $i = 1, \ldots, m$ indicates the cluster membership of the point $\mathbf{p}_i \in \mathbb{R}^m$. So, $\mathbf{X}_{ij} = 1/\sqrt{s_j}$ if and only if data point $\mathbf{p}_i$ is in cluster S_j. Every row of $\mathbf{X}$ has exactly one non-zero element, corresponding to the cluster the data point belongs to. There are s_j non-zero elements in column j which indicates the data points belonging to cluster S_j. By slightly abusing notation, we define

$$\mathcal{F}(\mathbf{A}, \mathbf{X}) := \|\mathbf{A} - \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2.$$

Hence, for any cluster indicator matrix $\mathbf{X}$, the following identities hold

$$\begin{aligned}\mathcal{F}(\mathbf{A}, \mathbf{X}) &= \sum_{i=1}^{m} \|\mathbf{p}_i^{\mathrm{T}} - \mathbf{X}_i\mathbf{X}^{\mathrm{T}}\mathbf{A}\|_2^2 \\ &= \sum_{i=1}^{m} \|\mathbf{p}_i^{\mathrm{T}} - \boldsymbol{\mu}(\mathbf{p}_i)^{\mathrm{T}}\|_2^2 \\ &= \mathcal{F}(\mathcal{P}, \mathcal{S}),\end{aligned}$$

where we define $\mathbf{X}_i$ as the ith row of $\mathbf{X}$ and we have used the identity $\mathbf{X}_i\mathbf{X}^{\mathrm{T}}\mathbf{A} = \boldsymbol{\mu}(p_i)^{\mathrm{T}}$, for $i = 1, ..., m$. This identity is true because $\mathbf{X}^{\mathrm{T}}\mathbf{A}$ is a matrix whose row j is $\sqrt{s_j}\boldsymbol{\mu}_j$, proportional to the centroid of the jth cluster; now, $\mathbf{X}_i$ picks the row corresponding to its non-zero element, i.e., the cluster corresponding to point i, and scales it by $1/\sqrt{s_j}$. In the above, $\mu(\mathbf{p}_i) \in \mathbb{R}^m$ denotes the centroid of the cluster of which the point $\mathbf{p}_i$ belongs to. Using this formulation, the goal of k-means is to find $\mathbf{X}$ which minimizes $\|\mathbf{A} - \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2$.

To evaluate the quality of different clusterings, we will use the k-means objective function. Given some clustering $\hat{\mathbf{X}}$, we are interested in the ratio $\mathcal{F}(\mathbf{A}, \hat{\mathbf{X}})/\mathcal{F}(\mathbf{A}, \mathbf{X}_{\mathrm{opt}})$, where $\mathbf{X}_{\mathrm{opt}}$ is an optimal clustering of $\mathbf{A}$. The choice of evaluating a clustering under this framework is not new. In fact, [3], [6]–[10] provide results (other than dimensionality reduction methods) along the same lines. Below, we give the definition of the k-means problem.

Definition 1 [The k-Means Clustering Problem]: Given $\mathbf{A} \in \mathbb{R}^{m\times n}$ (representing m data points – rows – described with respect to n features – columns) and a positive integer k denoting the number of clusters, find the indicator matrix $\mathbf{X}_{\mathrm{opt}} \in \mathbb{R}^{m\times k}$ which satisfies,

$$\mathbf{X}_{\mathrm{opt}} = \arg\min_{\mathbf{X}\in\mathcal{X}} \|\mathbf{A} - \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2.$$

The optimal value of the k-means clustering objective is

$$\begin{aligned}\mathcal{F}(\mathbf{A}, \mathbf{X}_{\mathrm{opt}}) &= \min_{\mathbf{X}\in\mathcal{X}} \|\mathbf{A} - \mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2 \\ &= \|\mathbf{A} - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2 \\ &= \mathrm{F}_{\mathrm{opt}}.\end{aligned}$$

In the above, $\mathcal{X}$ denotes the set of all $m \times k$ indicator matrices $\mathbf{X}$.

Next, we formalize the notation of a "k-means approximation algorithm".

Definition 2 [k-Means Approximation Algorithm]: An algorithm is called a "γ-approximation" for the k-means clustering problem ($\gamma \geq 1$) if it takes inputs the dataset $\mathbf{A} \in \mathbb{R}^{m\times n}$ and the number of clusters k, and returns an indicator matrix

$\mathbf{X}_\gamma \in \mathbb{R}^{m\times k}$ *such that w.p. at least* $1-\delta_\gamma$,

$$\begin{aligned}\|\mathbf{A}-\mathbf{X}_\gamma\mathbf{X}_\gamma^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2 &\le \gamma \min_{\mathbf{X}\in\mathcal{X}}\|\mathbf{A}-\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2\\ &= \gamma\cdot\mathcal{F}(\mathbf{A},\mathbf{X}_{\mathrm{opt}})\\ &= \gamma\cdot \mathrm{F}_{\mathrm{opt}}.\end{aligned}$$

An example of such an approximation algorithm for k-means is in [7] with $\gamma = 1+\varepsilon$ $(0<\varepsilon<1)$, and δ_γ a constant in $(0,1)$. The corresponding running time is $O(mn\cdot 2^{(k/\varepsilon)^{O(1)}})$.

Combining this algorithm (with $\gamma = 1+\varepsilon$) with, for example, our dimensionality reduction method in Section V, would result in an algorithm that preserves the clustering within a factor of $2+\varepsilon$, for any $\varepsilon\in(0,1/3)$, and runs in total time $O(mn\lceil\varepsilon^{-2}k/\log(n)\rceil + kn2^{(k/\varepsilon)^{O(1)}}/\varepsilon^2)$. Compare this with the complexity of running this algorithm on the high dimensional data and notice that reducing the dimension from n to $O(k/\varepsilon^2)$ leads to a considerably faster algorithm. In practice though, the Lloyd algorithm [2], [3] is very popular and although it does not admit a worst case theoretical analysis, it empirically does well. We thus employ the Lloyd algorithm for our experimental evaluation of our algorithms in Section VII. Note that, after using, for example, the dimensionality reduction method in Section V, the cost of the Lloyd heuristic is only $O(mk^2/\varepsilon^2)$ per iteration. This should be compared to the cost of $O(kmn)$ per iteration if applied on the original high dimensional data. Similar run time improvements arise if one uses the other dimension reduction algorithms proposed in this work.

B. Our Approach

The key insight of our work is to view the k-means problem from the above linear algebraic perspective. In this setting, the data points are rows in a matrix $\mathbf{A}$ and feature selection corresponds to selection of a subset of columns from $\mathbf{A}$. Also, feature extraction can be viewed as the construction of a matrix $\mathbf{C}$ which contains the constructed features. Our feature extraction algorithms are linear, i.e., the matrix $\mathbf{C}$ is of the form $\mathbf{C}=\mathbf{AD}$, for some matrix $\mathbf{D}$; so, the columns in $\mathbf{C}$ are linear combinations of the columns of $\mathbf{A}$, i.e., the new features are linear combinations of the original features.

Our work is inspired by the SVD feature extraction algorithm of [12], which also viewed the k-means problem from a linear algebraic perspective. The main message of the result of [12] (see the algorithm and the analysis in Section 2 in [12]) is that any matrix $\mathbf{C}$ which can be used to approximate the matrix $\mathbf{A}$ in some low-rank matrix approximation sense can also be used for dimensionality reduction in k-means clustering. We will now present a short proof of this result to better understand its implications in our dimensionality reduction algorithms.

Given $\mathbf{A}$ and k, the main algorithm of [12] constructs $\mathbf{C}=\mathbf{AV}_k$, where $\mathbf{V}_k$ contains the top k right singular vectors of $\mathbf{A}$. Let $\mathbf{X}_{opt}$ and $\hat{\mathbf{X}}_{opt}$ be the cluster indicator matrices that corresponds to the optimum partition corresponding to the rows of $\mathbf{A}$ and the rows of $\mathbf{C}$, respectively. In our setting for dimensionality reduction, we compare $\mathcal{F}(\mathbf{A},\hat{\mathbf{X}}_{opt})$ to $\mathcal{F}(\mathbf{A},\mathbf{X}_{opt})$. From the SVD of $\mathbf{A}$, consider

$$\mathbf{A}=\underbrace{\mathbf{AV}_k\mathbf{V}_k^{\mathrm{T}}}_{\mathbf{A}_k}+\underbrace{\mathbf{A}-\mathbf{AV}_k\mathbf{V}_k^{\mathrm{T}}}_{\mathbf{A}_{\rho-k}}.$$

Also, notice that for any cluster indicator matrix $\hat{\mathbf{X}}_{opt}$

$$\left(\left(\mathbf{I}_m-\hat{\mathbf{X}}_{opt}\hat{\mathbf{X}}_{opt}^{\mathrm{T}}\right)\mathbf{A}_k\right)\left(\left(\mathbf{I}_m-\hat{\mathbf{X}}_{opt}\hat{\mathbf{X}}_{opt}^{\mathrm{T}}\right)\mathbf{A}_{\rho-k}\right)^{\mathrm{T}}=\mathbf{0}_{m\times m},$$

because

$$\mathbf{A}_k\mathbf{A}_{\rho-k}^{\mathrm{T}}=\mathbf{0}_{m\times m}.$$

Combining these two steps and by orthogonality, it follows that

$$\begin{aligned}\|\mathbf{A}-\hat{\mathbf{X}}_{opt}\hat{\mathbf{X}}_{opt}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2 &= \underbrace{\|(\mathbf{I}_m-\hat{\mathbf{X}}_{opt}\hat{\mathbf{X}}_{opt}^{\mathrm{T}})\mathbf{A}_k\|_{\mathrm{F}}^2}_{\theta_\alpha^2}\\ &\quad+\underbrace{\|(\mathbf{I}_m-\hat{\mathbf{X}}_{opt}\hat{\mathbf{X}}_{opt}^{\mathrm{T}})\mathbf{A}_{\rho-k}\|_{\mathrm{F}}^2}_{\theta_\beta^2}.\end{aligned}$$

We now bound the second term of the later equation. $\mathbf{I}_m-\hat{\mathbf{X}}_{opt}\hat{\mathbf{X}}_{opt}^{\mathrm{T}}$ is a projection matrix, so it can be dropped without increasing the Frobenius norm. Hence, by using this and the fact that $\mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{A}$ has rank at most k:

$$\theta_\beta^2 \le \|\mathbf{A}_{\rho-k}\|_{\mathrm{F}}^2 = \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2 \le \|\mathbf{A}-\mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2.$$

From similar manipulations combined with the optimality of $\hat{\mathbf{X}}_{opt}$, it follows that

$$\theta_\alpha^2 \le \|\mathbf{A}-\mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2.$$

Therefore, we conclude that

$$\mathcal{F}(\mathbf{A},\hat{\mathbf{X}}_{opt})\le 2\mathcal{F}(\mathbf{A},\mathbf{X}_{opt}).$$

The key insight in this approach is that $\mathbf{A}_k=\mathbf{AV}_k\mathbf{V}_k^{\mathrm{T}}=\mathbf{C}\cdot\mathbf{H}$ (with $\mathbf{H}=\mathbf{V}_k^{\mathrm{T}}$) and $\mathbf{A}-\mathbf{CH}=\mathbf{A}_{\rho-k}$, which is the best rank k approximation of $\mathbf{A}$ in the Frobenius norm (see Section III for useful notation).

In all three methods of our work, we will construct matrices $\mathbf{C}=\mathbf{AD}$, for three different matrices $\mathbf{D}$, such that $\mathbf{C}\cdot\mathbf{H}$, for an appropriate $\mathbf{H}$, is a good approximation to $\mathbf{A}$ with respect to the Frobenius norm, i.e., $\|\mathbf{A}-\mathbf{C}\cdot\mathbf{H}\|_{\mathrm{F}}^2$ is roughly equal to $\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2$, where $\mathbf{A}_k$ is the best rank k matrix from the SVD of $\mathbf{A}$. Then, replicating the above proof gives our main theorems. Notice that the above approach is a 2-approximation because $\mathbf{A}_k=\mathbf{C}\cdot\mathbf{H}$ is the best rank k approximation to $\mathbf{A}$; our algorithms will give a slightly worse error because our matrix $\mathbf{C}\cdot\mathbf{H}$ give an approximation which is slightly worse than the best rank k approximation.

III. Preliminaries

A. Basic Notation

We use $\mathbf{A},\mathbf{B},\ldots$ to denote matrices; $\mathbf{a},\mathbf{p},\ldots$ to denote column vectors. $\mathbf{I}_n$ is the $n\times n$ identity matrix; $\mathbf{0}_{m\times n}$ is the $m\times n$ matrix of zeros; $\mathbf{A}_{(i)}$ is the i-th row of $\mathbf{A}$; $\mathbf{A}^{(j)}$ is the j-th column of $\mathbf{A}$; and, $\mathbf{A}_{ij}$ denotes the (i,j)-th element of $\mathbf{A}$. We use $\mathbb{E}\,Y$ to take the expectation of a random variable Y and $\mathbb{P}(\mathcal{E})$ to take the probability of a probabilistic event $\mathcal{E}$.

We abbreviate "independent identically distributed" to "i.i.d" and "with probability" to "w.p".

B. Matrix Norms

We use the Frobenius and the spectral matrix norms: $\|\mathbf{A}\|_\mathrm{F} = \sqrt{\sum_{i,j} \mathbf{A}_{ij}^2}$ and $\|\mathbf{A}\|_2 = \max_{\mathbf{x}:\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2$, respectively (for a matrix $\mathbf{A}$). For any $\mathbf{A}$,$\mathbf{B}$: $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_\mathrm{F}$, $\|\mathbf{AB}\|_\mathrm{F} \leq \|\mathbf{A}\|_\mathrm{F}\|\mathbf{B}\|_2$, and $\|\mathbf{AB}\|_\mathrm{F} \leq \|\mathbf{A}\|_2\|\mathbf{B}\|_\mathrm{F}$. The latter two properties are stronger versions of the standard submultiplicativity property: $\|\mathbf{AB}\|_\mathrm{F} \leq \|\mathbf{A}\|_\mathrm{F}\|\mathbf{B}\|_\mathrm{F}$. We will refer to these versions as spectral submultiplicativity. Finally, the triangle inequality of matrix norms indicates that $\|\mathbf{A}+\mathbf{B}\|_\mathrm{F} \leq \|\mathbf{A}\|_\mathrm{F} + \|\mathbf{B}\|_\mathrm{F}$.

Lemma 1 (Matrix Pythagorean Theorem): Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$ *satisfy* $\mathbf{X}\mathbf{Y}^\mathrm{T} = \mathbf{0}_{m\times m}$. *Then,*

$$\|\mathbf{X}+\mathbf{Y}\|_\mathrm{F}^2 = \|\mathbf{X}\|_\mathrm{F}^2 + \|\mathbf{Y}\|_\mathrm{F}^2.$$

Proof:

$$\begin{aligned}
\|\mathbf{X}+\mathbf{Y}\|_\mathrm{F}^2 &= \mathbf{Tr}\left((\mathbf{X}+\mathbf{Y})(\mathbf{X}+\mathbf{Y})^\mathrm{T}\right) \\
&= \mathbf{Tr}\left(\mathbf{X}\mathbf{X}^\mathrm{T} + \mathbf{X}\mathbf{Y}^\mathrm{T} + \mathbf{Y}\mathbf{X}^\mathrm{T} + \mathbf{Y}\mathbf{Y}^\mathrm{T}\right) \\
&= \mathbf{Tr}\left(\mathbf{X}\mathbf{X}^\mathrm{T} + \mathbf{0}_{m\times m} + \mathbf{0}_{m\times m} + \mathbf{Y}\mathbf{Y}^\mathrm{T}\right) \\
&= \mathbf{Tr}\left(\mathbf{X}^\mathrm{T}\mathbf{X}\right) + \mathbf{Tr}\left(\mathbf{Y}\mathbf{Y}^\mathrm{T}\right) \\
&= \|\mathbf{X}\|_\mathrm{F}^2 + \|\mathbf{Y}\|_\mathrm{F}^2.
\end{aligned}$$

■

This matrix form of the Pythagorean theorem is the starting point for the proofs of the three main theorems presented in this work. The idea to use the Matrix Pythagorean theorem to analyze a dimensionality reduction method for k-means was initially introduced in [12] and it turns to be very useful to prove our results as well.

C. Singular Value Decomposition

The SVD of $\mathbf{A} \in \mathbb{R}^{m\times n}$ of rank $\rho \leq \min\{m,n\}$ is $\mathbf{A} = \mathbf{U}_\mathbf{A}\Sigma_\mathbf{A}\mathbf{V}_\mathbf{A}^\mathrm{T}$, with $\mathbf{U}_\mathbf{A} \in \mathbb{R}^{m\times\rho}$, $\Sigma_\mathbf{A} \in \mathbb{R}^{\rho\times\rho}$, and $\mathbf{V}_\mathbf{A} \in \mathbb{R}^{n\times\rho}$. In some more details, the SVD of $\mathbf{A}$ is:

$$\mathbf{A} = \underbrace{\begin{pmatrix}\mathbf{U}_k & \mathbf{U}_{\rho-k}\end{pmatrix}}_{\mathbf{U}_\mathbf{A}\in\mathbb{R}^{m\times\rho}} \underbrace{\begin{pmatrix}\Sigma_k & \mathbf{0} \\ \mathbf{0} & \Sigma_{\rho-k}\end{pmatrix}}_{\Sigma_\mathbf{A}\in\mathbb{R}^{\rho\times\rho}} \underbrace{\begin{pmatrix}\mathbf{V}_k^\mathrm{T} \\ \mathbf{V}_{\rho-k}^\mathrm{T}\end{pmatrix}}_{\mathbf{V}_\mathbf{A}^\mathrm{T}\in\mathbb{R}^{\rho\times n}},$$

with singular values $\sigma_1 \geq \ldots \geq \sigma_k \geq \sigma_{k+1} \geq \ldots \geq \sigma_\rho > 0$. We will use $\sigma_i(\mathbf{A})$ to denote the i-th singular value of $\mathbf{A}$ when the matrix is not clear from the context. The matrices $\mathbf{U}_k \in \mathbb{R}^{m\times k}$ and $\mathbf{U}_{\rho-k} \in \mathbb{R}^{m\times(\rho-k)}$ contain the left singular vectors of $\mathbf{A}$; and, similarly, the matrices $\mathbf{V}_k \in \mathbb{R}^{n\times k}$ and $\mathbf{V}_{\rho-k} \in \mathbb{R}^{n\times(\rho-k)}$ contain the right singular vectors. $\Sigma_k \in \mathbb{R}^{k\times k}$ and $\Sigma_{\rho-k} \in \mathbb{R}^{(\rho-k)\times(\rho-k)}$ contain the singular values of $\mathbf{A}$. It is well-known that $\mathbf{A}_k = \mathbf{U}_k\Sigma_k\mathbf{V}_k^\mathrm{T} = \mathbf{A}\mathbf{V}_k\mathbf{V}_k^\mathrm{T} = \mathbf{U}_k\mathbf{U}_k^\mathrm{T}\mathbf{A}$ minimizes $\|\mathbf{A}-\mathbf{X}\|_\mathrm{F}$ over all matrices $\mathbf{X} \in \mathbb{R}^{m\times n}$ of rank at most $k \leq \rho$. We use $\mathbf{A}_{\rho-k} = \mathbf{A} - \mathbf{A}_k = \mathbf{U}_{\rho-k}\Sigma_{\rho-k}\mathbf{V}_{\rho-k}^\mathrm{T}$. Also, $\|\mathbf{A}\|_\mathrm{F} = \sqrt{\sum_{i=1}^\rho \sigma_i^2(\mathbf{A})}$ and $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$. The best rank k approximation to $\mathbf{A}$ also satisfies: $\|\mathbf{A}-\mathbf{A}_k\|_\mathrm{F} = \sqrt{\sum_{i=k+1}^\rho \sigma_i^2(\mathbf{A})}$.

D. Approximate Singular Value Decomposition

The exact SVD of $\mathbf{A}$ takes cubic time. In this work, to speed up certain algorithms, we will use fast approximate SVD. We quote a recent result from [19], but similar relative-error Frobenius norm SVD approximations can be found elsewhere; see, for example, [18].

Lemma 2: Given $\mathbf{A} \in \mathbb{R}^{m\times n}$ *of rank* ρ, *a target rank* $2 \leq k < \rho$, *and* $0 < \varepsilon < 1$, *there exists a randomized algorithm that computes a matrix* $\mathbf{Z} \in \mathbb{R}^{n\times k}$ *such that* $\mathbf{Z}^\mathrm{T}\mathbf{Z} = \mathbf{I}_k$, $\mathbf{E}\mathbf{Z} = \mathbf{0}_{m\times k}$ *(for* $\mathbf{E} = \mathbf{A} - \mathbf{A}\mathbf{Z}\mathbf{Z}^\mathrm{T} \in \mathbb{R}^{m\times n}$*), and*

$$\mathbb{E}\,\|\mathbf{E}\|_\mathrm{F}^2 \leq (1+\varepsilon)\,\|\mathbf{A}-\mathbf{A}_k\|_\mathrm{F}^2.$$

The proposed algorithm runs in $O(mnk/\varepsilon)$ *time. We use* $\mathbf{Z} = \mathit{FastFrobeniusSVD}(\mathbf{A}, k, \varepsilon)$ *to denote this algorithm.*

Notice that this lemma computes a rank-k matrix $\mathbf{A}\mathbf{Z}\mathbf{Z}^\mathrm{T}$ which, when is used to approximate $\mathbf{A}$, is almost as good - in expectation - as the rank-k matrix $\mathbf{A}_k$ from the SVD of $\mathbf{A}$. Since, $\mathbf{A}_k = \mathbf{A}\mathbf{V}_k\mathbf{V}_k^\mathrm{T}$, the matrix $\mathbf{Z}$ is essentially an approximation of the matrix $\mathbf{V}_k$ from the SVD of $\mathbf{A}$.

We now give the details of the algorithm. The algorithm takes as inputs a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ of rank ρ and an integer $2 \leq k < \rho$. Set $r = k + \lceil \frac{k}{\varepsilon} + 1 \rceil$ and construct $\mathbf{Z}$ with the following algorithm.

1: Generate an $n\times r$ standard Gaussian matrix $\mathbf{R}$ whose entries are i.i.d. $\mathcal{N}(0,1)$ variables.
2: $\mathbf{Y} = \mathbf{A}\mathbf{R} \in \mathbb{R}^{m\times r}$.
3: Orthonormalize the columns of $\mathbf{Y}$ to construct the matrix $\mathbf{Q} \in \mathbb{R}^{m\times r}$.
4: Let $\mathbf{Z} \in \mathbb{R}^{n\times k}$ be the top k right singular vectors of $\mathbf{Q}^\mathrm{T}\mathbf{A} \in \mathbb{R}^{r\times n}$.

E. Pseudo-Inverse

$\mathbf{A}^\dagger = \mathbf{V}_\mathbf{A}\Sigma_\mathbf{A}^{-1}\mathbf{U}_\mathbf{A}^\mathrm{T} \in \mathbb{R}^{n\times m}$ denotes the so-called Moore-Penrose pseudo-inverse of $\mathbf{A}$ (here $\Sigma_\mathbf{A}^{-1}$ is the inverse of $\Sigma_\mathbf{A}$), i.e., the unique $n\times m$ matrix satisfying all four properties: $\mathbf{A} = \mathbf{A}\mathbf{A}^\dagger\mathbf{A}$, $\mathbf{A}^\dagger\mathbf{A}\mathbf{A}^\dagger = \mathbf{A}^\dagger$, $(\mathbf{A}\mathbf{A}^\dagger)^\mathrm{T} = \mathbf{A}\mathbf{A}^\dagger$, and $(\mathbf{A}^\dagger\mathbf{A})^\mathrm{T} = \mathbf{A}^\dagger\mathbf{A}$. By the SVD of $\mathbf{A}$ and $\mathbf{A}^\dagger$, it is easy to verify that, for all $i = 1,\ldots,\rho = \mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{A}^\dagger)$: $\sigma_i(\mathbf{A}^\dagger) = 1/\sigma_{\rho-i+1}(\mathbf{A})$. Finally, for any $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times\ell}$: $(\mathbf{AB})^\dagger = \mathbf{B}^\dagger\mathbf{A}^\dagger$ if any one of the following three properties hold: (i) $\mathbf{A}^\mathrm{T}\mathbf{A} = \mathbf{I}_n$; (ii) $\mathbf{B}^\mathrm{T}\mathbf{B} = \mathbf{I}_\ell$; or, (iii) $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{B}) = n$.

F. Projection Matrices

We call $\mathbf{P} \in \mathbb{R}^{n\times n}$ a projection matrix if $\mathbf{P}^2 = \mathbf{P}$. For such a projection matrix and any $\mathbf{A}$: $\|\mathbf{P}\mathbf{A}\|_\mathrm{F} \leq \|\mathbf{A}\|_\mathrm{F}$. Also, if $\mathbf{P}$ is a projection matrix, then, $\mathbf{I}_n - \mathbf{P}$ is a projection matrix. So, for any matrix $\mathbf{A}$, both $\mathbf{A}\mathbf{A}^\dagger$ and $\mathbf{I}_n - \mathbf{A}\mathbf{A}^\dagger$ are projection matrices.

G. Markov's Inequality and the Union Bound

Markov's inequality can be stated as follows: Let Y be a random variable taking non-negative values with expectation $\mathbb{E}\,Y$. Then, for all $t > 0$, and with probability at least $1 - t^{-1}$, $Y \leq t \cdot \mathbb{E}\,Y$. We will also use the so-called union bound. Given a set of probabilistic events $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ holding with respective probabilities $p_1, p_2, \ldots, p_n$, the probability that all

events hold simultaneously (a.k.a., the probability of the union of those events) is upper bounded as: $\mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2 \ldots \cup \mathcal{E}_n) \leq \sum_{i=1}^{n} p_i$.

H. Randomized Sampling

1) Sampling and Rescaling Matrices: Let $\mathbf{A} = [\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(n)}] \in \mathbb{R}^{m \times n}$ and let $\mathbf{C} = [\mathbf{A}^{(i_1)}, \ldots, \mathbf{A}^{(i_r)}] \in \mathbb{R}^{m \times r}$ consist of $r < n$ columns of $\mathbf{A}$. Note that we can write $\mathbf{C} = \mathbf{A}\Omega$, where the *sampling matrix* is $\Omega = [\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_r}] \in \mathbb{R}^{n \times r}$ (here $\mathbf{e}_i$ are the standard basis vectors in $\mathbb{R}^n$). If $\mathbf{S} \in \mathbb{R}^{r \times r}$ is a diagonal *rescaling matrix* then $\mathbf{A}\Omega\mathbf{S}$ contains r rescaled columns of $\mathbf{A}$.

The following definition describes a simple randomized sampling procedure with replacement, which will be critical in our feature selection algorithm.

Definition 3 (Random Sampling With Replacement): Let $\mathbf{X} \in \mathbb{R}^{n \times k}$ with $n > k$ and let $\mathbf{X}_{(i)}$ denote the i-th row of $\mathbf{X}$ as a row vector. For all $i = 1, \ldots, n$, define the following set of sampling probabilities:

$$p_i = \frac{\|\mathbf{X}_{(i)}\|_2^2}{\|\mathbf{X}\|_{\mathrm{F}}^2},$$

and note that $\sum_{i=1}^{n} p_i = 1$. Let r be a positive integer and construct the sampling matrix $\Omega \in \mathbb{R}^{n \times r}$ and the rescaling matrix $\mathbf{S} \in \mathbb{R}^{r \times r}$ as follows: initially, $\Omega = \mathbf{0}_{n \times r}$ and $\mathbf{S} = \mathbf{0}_{r \times r}$; for $t = 1, \ldots, r$ pick an integer i_t from the set $\{1, 2, \ldots, n\}$ where the probability of picking i is equal to p_i; set $\Omega_{i_t t} = 1$ and $\mathbf{S}_{tt} = 1/\sqrt{r p_{i_t}}$. We denote this randomized sampling technique with replacement by

$$[\Omega, \mathbf{S}] = \mathit{RandomizedSampling}(\mathbf{X}, r).$$

Given $\mathbf{X}$ and r, it takes $O(nk)$ time to compute the probabilities and another $O(n + r)$ time to implement the sampling procedure via the technique in [20]. In total, this method requires $O(nk)$ time.

The next three lemmas present the effect of the above sampling procedure on certain spectral properties, e.g. singular values, of orthogonal matrices. The first two lemmas are known; short proofs are included for the sake of completeness. The third lemma follows easily from the first two results (a proof of the lemma is given for completeness as well). We remind the reader that $\sigma_i^2(\mathbf{X})$ denotes the ith singular value squared of the matrix $\mathbf{X}$.

Lemma 3 argues that sampling and rescaling a sufficiently large number of rows from an orthonormal matrix with the randomized procedure of Definition 3 results in a matrix with singular values close to the singular values of the original orthonormal matrix.

Lemma 3: Let $\mathbf{V} \in \mathbb{R}^{n \times k}$ with $n > k$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_k$. Let $0 < \delta < 1$, $4k \ln(2k/\delta) < r \leq n$, and $[\Omega, \mathbf{S}] = \mathit{RandomizedSampling}(\mathbf{V}, r)$. Then, for all $i = 1, \ldots, k$, w.p. at least $1 - \delta$,

$$1 - \sqrt{\frac{4k \ln(2k/\delta)}{r}} \leq \sigma_i^2(\mathbf{V}^{\mathrm{T}}\Omega\mathbf{S}) \leq 1 + \sqrt{\frac{4k \ln(2k/\delta)}{r}}.$$

Proof: This result was originally proven in [17]. We will leverage a more recent proof of this result that appeared in [21] and improves the original constants. More specifically, in [21, Th. 2], set $\mathbf{S} = \mathbf{I}$, $\beta = 1$, and replace ε as a function of r, β, and d to conclude the proof. ∎

Lemma 4 argues that sampling and rescaling columns from any matrix with the randomized procedure of Definition 3 results in a matrix with Frobenius norm squared close to the Frobenius norm squared of the original matrix. Intuitively, the subsampling of the columns does not affect much the Frobenius norm of the matrix.

Lemma 4: For any $r \geq 1$, $\mathbf{X} \in \mathbb{R}^{n \times k}$, and $\mathbf{Y} \in \mathbb{R}^{m \times n}$, let $[\Omega, \mathbf{S}] = \mathit{RandomizedSampling}(\mathbf{X}, r)$. Let δ be a parameter with $0 < \delta < 1$. Then, w.p. at least $1 - \delta$,

$$\|\mathbf{Y}\Omega\mathbf{S}\|_{\mathrm{F}}^2 \leq \frac{1}{\delta}\|\mathbf{Y}\|_{\mathrm{F}}^2.$$

Proof: Define the random variable $Y = \|\mathbf{Y}\Omega\mathbf{S}\|_{\mathrm{F}}^2 \geq 0$. Assume that the following equation is true: $\mathbb{E}\,\|\mathbf{Y}\Omega\mathbf{S}\|_{\mathrm{F}}^2 = \|\mathbf{Y}\|_{\mathrm{F}}^2$. Applying Markov's inequality with failure probability δ to this equation gives the bound in the lemma. All that remains to prove now is the above assumption. Let $\mathbf{B} = \mathbf{Y}\Omega\mathbf{S} \in \mathbb{R}^{m \times r}$, and for $t = 1, \ldots, r$, let $\mathbf{B}^{(t)}$ denotes the t-th column of $\mathbf{B} = \mathbf{Y}\Omega\mathbf{S}$. We manipulate the term $\mathbb{E}\,\|\mathbf{Y}\Omega\mathbf{S}\|_{\mathrm{F}}^2$ as follows,

$$\begin{aligned}
\mathbb{E}\,\|\mathbf{Y}\Omega\mathbf{S}\|_{\mathrm{F}}^2 &\overset{(a)}{=} \mathbb{E}\sum_{t=1}^{r}\|\mathbf{B}^{(t)}\|_2^2 \overset{(b)}{=} \sum_{t=1}^{r}\mathbb{E}\,\|\mathbf{B}^{(t)}\|_2^2 \\
&\overset{(c)}{=} \sum_{t=1}^{r}\sum_{j=1}^{n} p_j \frac{\|\mathbf{Y}^{(j)}\|_2^2}{r p_j} \\
&\overset{(d)}{=} \frac{1}{r}\sum_{t=1}^{r}\|\mathbf{Y}\|_{\mathrm{F}}^2 = \|\mathbf{Y}\|_{\mathrm{F}}^2
\end{aligned}$$

(a) follows by the definition of the Frobenius norm of $\mathbf{B}$. (b) follows by the linearity of expectation. (c) follows by our construction of $\Omega, \mathbf{S}$. (d) follows by the definition of the Frobenius norm of $\mathbf{Y}$. It is worth noting that the above manipulations hold for any set of probabilities since they cancel out in Equation (d). ∎

Notice that $\mathbf{X}$ does not appear in the bound; it is only used as an input to the RandomizedSampling. This means that for *any* set of probabilities, a sampling and rescaling matrix constructed in the way it is described in Definition 3 satisfies the bound in the lemma.

The next lemma shows the effect of sub-sampling in a low-rank approximation of the form $\mathbf{A} \approx \mathbf{A}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$, where $\mathbf{Z}$ is a tall-and-skinny orthonormal matrix. The sub-sampling here is done on the columns of $\mathbf{A}$ and the corresponding rows of $\mathbf{Z}$.

Lemma 5: Fix $\mathbf{A} \in \mathbb{R}^{m \times n}$, $k \geq 1$, $0 < \varepsilon < 1/3$, $0 < \delta < 1$, and $r = 4k \ln(2k/\delta)/\varepsilon^2$. Compute the $n \times k$ matrix $\mathbf{Z}$ of Lemma 2 such that $\mathbf{A} = \mathbf{A}\mathbf{Z}\mathbf{Z}^{\mathrm{T}} + \mathbf{E}$ and run $[\Omega, \mathbf{S}] = \mathit{RandomizedSampling}(\mathbf{Z}, r)$. Then, w.p. at least $1 - 3\delta$, there exists $\widetilde{\mathbf{E}} \in \mathbb{R}^{m \times n}$ such that

$$\mathbf{A}\mathbf{Z}\mathbf{Z}^{\mathrm{T}} = \mathbf{A}\Omega\mathbf{S}(\mathbf{Z}^{\mathrm{T}}\Omega\mathbf{S})^{\dagger}\mathbf{Z}^{\mathrm{T}} + \widetilde{\mathbf{E}},$$

and $\|\widetilde{\mathbf{E}}\|_{\mathrm{F}} \leq \frac{1.6\varepsilon}{\sqrt{\delta}}\|\mathbf{E}\|_{\mathrm{F}}$.

Proof: See Appendix. ∎

In words, given $\mathbf{A}$ and the rank parameter k, it is possible to construct two low rank matrices, $\mathbf{AZZ}^{\mathrm{T}}$ and $\mathbf{A\Omega S}(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\dagger}\mathbf{Z}^{\mathrm{T}}$ that are "close" to each other. Another way to view this result is that given the low-rank factorization $\mathbf{AZZ}^{\mathrm{T}}$ one can "compress" $\mathbf{A}$ and $\mathbf{Z}$ by means of the sampling and rescaling matrices $\mathbf{\Omega}$ and $\mathbf{S}$. The error from such a compression will be bounded by $\tilde{\mathbf{E}}$.

This result is useful in proving Theorem 1 because at some point of the proof (see Eqn. (3)) we need to switch from a rank r matrix ($\mathbf{A\Omega S}(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\dagger}\mathbf{Z}^{\mathrm{T}}$) to a rank k matrix ($\mathbf{AZZ}^{\mathrm{T}}$) and at the same time keep the bounds in the resulting inequality almost unchanged (they would change by the norm of the matrix $\widetilde{\mathbf{E}}$).

I. Random Projections

A classic result of [15] states that, for any $0 < \varepsilon < 1$, any set of m points in n dimensions (rows in $\mathbf{A} \in \mathbb{R}^{m\times n}$) can be linearly projected into

$$r_{\varepsilon} = O\left(\log(m)/\varepsilon^2\right)$$

dimensions while preserving all the pairwise Euclidean distances of the points within a multiplicative factor of $(1 \pm \varepsilon)$. More precisely, [15] showed the existence of a (random orthonormal) matrix $\mathbf{R} \in \mathbb{R}^{n\times r_{\varepsilon}}$ such that, for all $i, j = 1, \ldots, m$, and with high probability (over the randomness of the matrix $\mathbf{R}$),

$$\begin{aligned}(1-\varepsilon)\|\mathbf{A}_{(i)} - \mathbf{A}_{(j)}\|_2 &\leq \|\left(\mathbf{A}_{(i)} - \mathbf{A}_{(j)}\right)\mathbf{R}\|_2 \\ &\leq (1+\varepsilon)\|\mathbf{A}_{(i)} - \mathbf{A}_{(j)}\|_2.\end{aligned}$$

Subsequent research simplified the proof of [15] by showing that such a linear transformation can be generated using a random Gaussian matrix, i.e., a matrix $\mathbf{R} \in \mathbb{R}^{n\times r_{\varepsilon}}$ whose entries are i.i.d. Gaussian random variables with zero mean and variance $1/r$ [22]. Recently, [23] presented the so-called Fast Johnson-Lindenstrauss Transform which describes an $\mathbf{R} \in \mathbb{R}^{n\times r_{\varepsilon}}$ such that the product $\mathbf{AR}$ can be computed fast. In this paper, we will use a construction by [24], who proved that a rescaled random sign matrix, i.e., a matrix $\mathbf{R} \in \mathbb{R}^{n\times r_{\varepsilon}}$ whose entries have values $\{\pm 1/\sqrt{r}\}$ uniformly at random, satisfies the above equation. As we will see in detail in Section V, a recent result of [25] indicates that, if $\mathbf{R}$ is constructed as in [24], the product $\mathbf{AR}$ can be computed fast as well. We utilize such a random projection embedding in Section V. Here, we summarize some properties of such matrices that might be of independent interest. We have deferred the proofs of the following lemmata to the Appendix.

The first lemma argues that the Frobenius norm squared of any matrix $\mathbf{Y}$ and the Frobenius norm squared of $\mathbf{YR}$, where $\mathbf{R}$ is a scaled signed matrix, are "comparable". Lemma 6 is the analog of Lemma 4.

Lemma 6: Fix any $m \times n$ matrix $\mathbf{Y}$, fix $k > 1$ and $\varepsilon > 0$. Let $\mathbf{R} \in \mathbb{R}^{n\times r}$ be a rescaled random sign matrix constructed as described above with $r = c_0 k/\varepsilon^2$, where $c_0 \geq 100$. Then,

$$\mathbb{P}\left(\|\mathbf{YR}\|_{\mathrm{F}}^2 \geq (1+\varepsilon)\|\mathbf{Y}\|_{\mathrm{F}}^2\right) \leq 0.01.$$

The next lemma argues about the effect of scaled random signed matrices to the singular values of orthonormal matrices.

Lemma 7: Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ with rank ρ ($k < \rho$), $\mathbf{A}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^{\mathrm{T}}$, and $0 < \varepsilon < 1/3$. Let $\mathbf{R} \in \mathbb{R}^{n\times r}$ be a (rescaled) random sign matrix constructed as we described above with $r = c_0 k/\varepsilon^2$, where $c_0 \geq 3330$. The following hold (simultaneously) w.p. at least 0.97:

1) For all $i = 1, \ldots, k$:

$$1 - \varepsilon \leq \sigma_i^2(\mathbf{V}_k^{\mathrm{T}}\mathbf{R}) \leq 1 + \varepsilon.$$

2) There exists an $m \times n$ matrix $\widetilde{\mathbf{E}}$ such that

$$\mathbf{A}_k = \mathbf{AR}(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\dagger}\mathbf{V}_k^{\mathrm{T}} + \widetilde{\mathbf{E}},$$

and

$$\|\widetilde{\mathbf{E}}\|_{\mathrm{F}} \leq 3\varepsilon\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}.$$

The first statement of Lemma 7 is the analog of Lemma 3 while the second statement of Lemma 7 is the analog of Lemma 5. The results here replace the sampling and rescaling matrices $\mathbf{\Omega}, \mathbf{S}$ from Random Sampling (Definition 3) with the Random Projection matrix $\mathbf{R}$. It is worth noting that almost the same results can be achieved with $r = O(k/\varepsilon^2)$ random dimensions, while the corresponding lemmata for Random Sampling require at least $r = O(k\log k/\varepsilon^2)$ actual dimensions.

The second bound in the lemma is useful in proving Theorem 2. Specifically, in Eqn. (2) we need to replace the rank k matrix $\mathbf{A}_k$ with another matrix of rank k which is as close to $\mathbf{A}_k$ as possible. The second bound above provides precisely such a matrix $\mathbf{AR}(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\dagger}\mathbf{V}_k^{\mathrm{T}}$ with corresponding error $\widetilde{\mathbf{E}}$.

IV. Feature Selection With Randomized Sampling

Given $\mathbf{A}, k$, and $0 < \varepsilon < 1/3$, Algorithm 1 is our main algorithm for feature selection in k-means clustering. In a nutshell, construct the matrix $\mathbf{Z}$ with the (approximate) top-k right singular vectors of $\mathbf{A}$ and select

$$r = O(k\log(k)/\varepsilon^2)$$

columns from $\mathbf{Z}^{\mathrm{T}}$ with the randomized technique of Section III-A. One can replace the first step in Algorithm 1 with the exact SVD of $\mathbf{A}$ [16]. The result that is obtained from this approach is asymptotically the same as the one we will present in Theorem 1.[2] Working with $\mathbf{Z}$ though gives a considerably faster algorithm.

Theorem 1: Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and k be inputs of the k-means clustering problem. Let $\varepsilon \in (0, 1/3)$ and, by using Algorithm 1 in $O(mnk/\varepsilon + k\ln(k)/\varepsilon^2\log(k\ln(k)/\varepsilon))$ time construct features $\mathbf{C} \in \mathbb{R}^{m\times r}$ with $r = O(k\log(k)/\varepsilon^2)$. Run any γ-approximation k-means algorithm with failure probability δ_{γ} on $\mathbf{C}, k$ and construct $\mathbf{X}_{\tilde{\gamma}}$. Then, w.p.

[2]The main theorem of [16] states a $(1 + (1+\varepsilon)\gamma)$-approximation bound but the corresponding proof has a bug, which is fixable and leads to a $(1 + (2+\varepsilon)\gamma)$-approximation bound. One can replicate the corresponding (fixable) proof in [16] by replacing $\mathbf{Z} = \mathbf{V}_k$ in the proof of Theorem 1 of our work.

at least $0.2 - \delta_\gamma$,

$$\|\mathbf{A} - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2 \leq (1 + (2+\varepsilon)\gamma)\,\|\mathbf{A} - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2.$$

In words, given any set of points in some n-dimensional space and the number of clusters k, it suffices to select roughly $O(k\log k)$ actual features from the given points and then run some k-means algorithm on this subset of the input. The theorem formally argues that the clustering it would be obtained in the low-dimensional space will be close to the clustering it would have been obtained after running the k-means method in the original high-dimensional data. We also state the result of the theorem in the notation we introduced in Section I,

$$\mathcal{F}(\mathcal{P}, \mathcal{S}_{\tilde{\gamma}}) \leq (1 + (2+\varepsilon)\gamma)\,\mathcal{F}(\mathcal{P}, \mathcal{S}_{opt}).$$

Here, $\mathcal{S}_{\tilde{\gamma}}$ is the partition obtained after running the γ-approximation k-means algorithm on the low-dimensional space. The approximation factor is $(1 + (2+\varepsilon)\gamma)$. The term $\gamma > 1$ is due to the fact that the k-means method that we run in the low-dimensional space does not recover the optimal k-means partition. The other factor $2 + \varepsilon$ is due to the fact that we run k-means in the low-dimensional space.

Proof (of Theorem 1): We start by manipulating the term $\|\mathbf{A} - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2$. Notice that

$$\mathbf{A} = \mathbf{A}\mathbf{Z}\mathbf{Z}^{\mathrm{T}} + \mathbf{E},$$

(from Lemma 2). Also,

$$\left(\left(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\right)\mathbf{A}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\right)\left(\left(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\right)\mathbf{E}\right)^{\mathrm{T}} = \mathbf{0}_{m\times m},$$

because

$$\mathbf{Z}^{\mathrm{T}}\mathbf{E}^{\mathrm{T}} = \mathbf{0}_{k\times m},$$

by construction. Now, using Matrix Pythagoras (see Lemma 1),

$$\|\mathbf{A} - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2 = \underbrace{\|(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}})\mathbf{A}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\|_{\mathrm{F}}^2}_{\theta_1^2} + \underbrace{\|(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}})\mathbf{E}\|_{\mathrm{F}}^2}_{\theta_2^2}. \quad (1)$$

We first bound the second term of Eqn. (1). Since $\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}$ is a projection matrix, it can be dropped without increasing the Frobenius norm (see Section III). Applying Markov's inequality on the equation of Lemma 2, we obtain that w.p. 0.99,

$$\|\mathbf{E}\|_{\mathrm{F}}^2 \leq (1 + 100\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}^2. \quad (2)$$

(See the Appendix for a short proof of this statement.) Note also that $\mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{A}$ has rank at most k; so, from the optimality of the SVD, overall,

$$\begin{aligned}\theta_2^2 &\leq (1 + 100\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}^2\\ &\leq (1 + 100\varepsilon)\|\mathbf{A} - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2\\ &= (1 + 100\varepsilon)\mathrm{F}_{\mathrm{opt}}.\end{aligned}$$

We now bound the first term in Eqn. (1),

$$\begin{aligned}\theta_1 &\leq \|(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}})\mathbf{A}\Omega\mathbf{S}(\mathbf{Z}^{\mathrm{T}}\Omega\mathbf{S})^{\dagger}\mathbf{Z}^{\mathrm{T}}\|_{\mathrm{F}} + \|\widetilde{\mathbf{E}}\|_{\mathrm{F}} &(3)\\ &\leq \|(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}})\mathbf{A}\Omega\mathbf{S}\|_{\mathrm{F}}\|(\mathbf{Z}^{\mathrm{T}}\Omega\mathbf{S})^{\dagger}\|_2 + \|\widetilde{\mathbf{E}}\|_{\mathrm{F}} &(4)\\ &\leq \sqrt{\gamma}\,\|(\mathbf{I}_m - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}})\mathbf{A}\Omega\mathbf{S}\|_{\mathrm{F}}\|(\mathbf{Z}^{\mathrm{T}}\Omega\mathbf{S})^{\dagger}\|_2 + \|\widetilde{\mathbf{E}}\|_{\mathrm{F}} &(5)\end{aligned}$$

In Eqn. (3), we used Lemma 5 (for an unspecified failure probability δ; also, $\widetilde{\mathbf{E}} \in \mathbb{R}^{m\times n}$ is from that lemma), the triangle inequality, and the fact that $\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}$ is a projection matrix and can be dropped without increasing the Frobenius norm. In Eqn. (4), we used spectral submultiplicativity and the fact that $\mathbf{Z}^{\mathrm{T}}$ can be dropped without changing the spectral norm. In Eqn. (5), we replaced $\mathbf{X}_{\tilde{\gamma}}$ by $\mathbf{X}_{\mathrm{opt}}$ and the factor $\sqrt{\gamma}$ appeared in the first term. To better understand this step, notice that $\mathbf{X}_{\tilde{\gamma}}$ gives a γ-approximation to the optimal k-means clustering of $\mathbf{C} = \mathbf{A}\Omega\mathbf{S}$, so any other $m \times k$ indicator matrix (e.g. $\mathbf{X}_{\mathrm{opt}}$) satisfies,

$$\begin{aligned}\|\left(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\right)\mathbf{A}\Omega\mathbf{S}\|_{\mathrm{F}}^2 &\leq \gamma \min_{\mathbf{X}\in\mathcal{X}}\|(\mathbf{I}_m - \mathbf{X}\mathbf{X}^{\mathrm{T}})\mathbf{A}\Omega\mathbf{S}\|_{\mathrm{F}}^2\\ &\leq \gamma\,\|\left(\mathbf{I}_m - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\right)\mathbf{A}\Omega\mathbf{S}\|_{\mathrm{F}}^2.\end{aligned}$$

By using Lemma 4 with $\delta = 3/4$ and Lemma 3 (for an unspecified failure probability δ),

$$\|(\mathbf{I}_m - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}})\mathbf{A}\Omega\mathbf{S}\|_{\mathrm{F}}\|(\mathbf{Z}^{\mathrm{T}}\Omega\mathbf{S})^{\dagger}\|_2 \leq \sqrt{\frac{4}{3-3\varepsilon}}\mathrm{F}_{\mathrm{opt}}.$$

We are now in position to bound θ_1. In Lemmas 5 and 3, let $\delta = 0.01$. Assuming $1 \leq \gamma$,

$$\begin{aligned}\theta_1 &\leq \left(\sqrt{\frac{4}{3-3\varepsilon}} + \frac{1.6\varepsilon\sqrt{1+100\varepsilon}}{\sqrt{0.01}}\right)\sqrt{\gamma}\sqrt{\mathrm{F}_{\mathrm{opt}}}\\ &\leq \left(\sqrt{2} + 94\varepsilon\right)\sqrt{\gamma}\sqrt{\mathrm{F}_{\mathrm{opt}}}.\end{aligned}$$

The last inequality follows from our choice of $\varepsilon < 1/3$ and elementary algebra. Taking squares on both sides,

$$\theta_1^2 \leq \left(\sqrt{2} + 94\varepsilon\right)^2\gamma\,\mathrm{F}_{\mathrm{opt}} \leq (2 + 3900\varepsilon)\gamma\,\mathrm{F}_{\mathrm{opt}}.$$

Overall (assuming $1 \leq \gamma$),

$$\begin{aligned}\|\mathbf{A} - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2 &\leq \theta_1^2 + \theta_2^2\\ &\leq (2 + 3900\varepsilon)\gamma\,\mathrm{F}_{\mathrm{opt}} + (1 + 100\varepsilon)\mathrm{F}_{\mathrm{opt}}\\ &\leq \mathrm{F}_{\mathrm{opt}} + (2 + 4\cdot 10^3\varepsilon)\gamma\,\mathrm{F}_{\mathrm{opt}}.\end{aligned}$$

Rescaling ε accordingly ($c_1 = 16\cdot 10^6$) gives the bound in the Theorem. The failure probability follows by a union bound on Lemma 4 (with $\delta = 3/4$), Lemma 5 (with $\delta = 0.01$), Lemma 3 (with $\delta = 0.01$), Lemma 2 (followed by Markov's inequality with $\delta = 0.01$), and Definition 2 (with failure probability δ_γ). Indeed, $0.75 + 3\cdot 0.01 + 0.01 + 0.01 + \delta_\gamma = 0.8 + \delta_\gamma$ is the overall failure probability, hence the bound in the theorem holds w.p. $0.2 - \delta_\gamma$. ■

V. Feature Extraction With Random Projections

We prove that any set of m points in n dimensions (rows in a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$) can be projected into $r = O(k/\varepsilon^2)$ dimensions in $O(mn\lceil\varepsilon^{-2}k/\log(n)\rceil)$ time

Algorithm 2 Randomized Feature Extraction for k-Means Clustering

Input: Dataset $\mathbf{A} \in \mathbb{R}^{m\times n}$, number of clusters k, and $0 < \varepsilon < \frac{1}{3}$.
Output: $\mathbf{C} \in \mathbb{R}^{m\times r}$ with $r = O(k/\varepsilon^2)$ artificial features.

1: Set $r = c_2 \cdot k/\varepsilon^2$, for a sufficiently large constant c_2 (see proof).
2: Compute a random $n \times r$ matrix $\mathbf{R}$ as follows. For all $i = 1, \ldots, n,\ j = 1, \ldots, r$ (i.i.d.)

$$\mathbf{R}_{ij} = \begin{cases} +1/\sqrt{r}, \text{ w.p. } 1/2, \\ -1/\sqrt{r}, \text{ w.p. } 1/2. \end{cases}$$

3: Compute $\mathbf{C} = \mathbf{AR}$ with the Mailman Algorithm (see text).
4: Return $\mathbf{C} \in \mathbb{R}^{m\times r}$.

such that, with constant probability, the objective value of the optimal k-partition of the points is preserved within a factor of $2+\varepsilon$. The projection is done by post-multiplying $\mathbf{A}$ with an $n \times r$ random matrix $\mathbf{R}$ having entries $+1/\sqrt{r}$ or $-1/\sqrt{r}$ with equal probability.

The algorithm needs $O(mk/\varepsilon^2)$ time to generate $\mathbf{R}$; then, the product $\mathbf{AR}$ can be naively computed in $O(mnk/\varepsilon^2)$. However, one can employ the so-called mailman algorithm for matrix multiplication [25] and compute the product $\mathbf{AR}$ in $O(mn\lceil\varepsilon^{-2}k/\log(n)\rceil)$. Indeed, the mailman algorithm computes (after preprocessing) a matrix-vector product of any n-dimensional vector (row of $\mathbf{A}$) with an $n \times \log(n)$ sign matrix in $O(n)$ time. Reading the input $n \times \log n$ sign matrix requires $O(n \log n)$ time. However, in our case we only consider multiplication with a random sign matrix, therefore we can avoid the preprocessing step by directly computing a random correspondence matrix as discussed in [25, Preprocessing Section]. By partitioning the columns of our $n \times r$ matrix $\mathbf{R}$ into $\lceil r/\log(n)\rceil$ blocks, the desired running time follows.

Theorem 2 is our quality-of-approximation result regarding the clustering that can be obtained with the features returned from Algorithm 2. Notice that if $\gamma = 1$, the distortion is at most $2+\varepsilon$, as advertised in Table I. If the γ-approximation algorithm is [7] the overall approximation factor would be $(1+(1+\varepsilon)^2) = 2 + O(\varepsilon)$ with running time of the order $O(mn\lceil\varepsilon^{-2}k/\log(n)\rceil + 2^{(k/\varepsilon)^{O(1)}}mk/\varepsilon^2)$.

Theorem 2: Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and k be the inputs of the k-means clustering problem. Let $\varepsilon \in (0, 1/3)$ and construct features $\mathbf{C} \in \mathbb{R}^{m\times r}$ with $r = O(k/\varepsilon^2)$ by using Algorithm 2 in $O(mn\lceil\varepsilon^{-2}k/\log(n)\rceil)$ time. Run any γ-approximation k-means algorithm with failure probability δ_γ on $\mathbf{C}, k$ and construct $\mathbf{X}_{\tilde{\gamma}}$. Then, w.p. at least $0.96 - \delta_\gamma$,

$$\|\mathbf{A} - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2 \leq (1 + (1+\varepsilon)\gamma)\,\|\mathbf{A} - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2.$$

In words, given any set of points in some n-dimensional space and the number of clusters k, it suffices to create (via random projections) roughly $O(k)$ new features and then run some k-means algorithm on this new input. The theorem formally argues that the clustering it would be obtained in the low-dimensional space will be close to the clustering it would have been obtained after running the k-means method in the original high-dimensional data. We also state the result of the theorem in the notation we introduced in Section I,

$$\mathcal{F}(\mathcal{P}, \mathcal{S}_{\tilde{\gamma}}) \leq (1 + (1+\varepsilon)\gamma)\,\mathcal{F}(\mathcal{P}, \mathcal{S}_{opt}).$$

Here, $\mathcal{S}_{\tilde{\gamma}}$ is the partition obtained after running the γ-approximation k-means algorithm on the low-dimensional space. The approximation factor is $(1+(1+\varepsilon)\gamma)$. The term $\gamma > 1$ is due to the fact that the k-means method that we run in the low-dimensional space does not recover the optimal k-means partition. The other factor $1+\varepsilon$ is due to the fact that we run k-means in the low-dimensional space.

Proof (of Theorem 2): We start by manipulating the term $\|\mathbf{A} - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2$. Notice that $\mathbf{A} = \mathbf{A}_k + \mathbf{A}_{\rho-k}$. Also, $\left(\left(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\right)\mathbf{A}_k\right)\left(\left(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\right)\mathbf{A}_{\rho-k}\right)^{\mathrm{T}} = \mathbf{0}_{m\times m}$, because $\mathbf{A}_k\mathbf{A}_{\rho-k}^{\mathrm{T}} = \mathbf{0}_{m\times m}$, by the orthogonality of the corresponding subspaces. Now, using Lemma 1,

$$\|\mathbf{A} - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2 = \underbrace{\|(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}})\mathbf{A}_k\|_{\mathrm{F}}^2}_{\theta_3^2} + \underbrace{\|(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}})\mathbf{A}_{\rho-k}\|_{\mathrm{F}}^2}_{\theta_4^2}. \quad (6)$$

We first bound the second term of Eqn. (6). Since $\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}$ is a projection matrix, it can be dropped without increasing the Frobenius norm. So, by using this and the fact that $\mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{A}$ has rank at most k,

$$\theta_4^2 \leq \|\mathbf{A}_{\rho-k}\|_{\mathrm{F}}^2 = \|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}^2 \leq \|\mathbf{A} - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2. \quad (7)$$

We now bound the first term of Eqn. (6),

$$\begin{aligned}
\theta_3 &\leq \|(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}})\mathbf{AR}(\mathbf{V}_k\mathbf{R})^{\dagger}\mathbf{V}_k^{\mathrm{T}}\|_{\mathrm{F}} + \|\widetilde{\mathbf{E}}\|_{\mathrm{F}} && (8)\\
&\leq \|(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}})\mathbf{AR}\|_{\mathrm{F}}\|(\mathbf{V}_k\mathbf{R})^{\dagger}\|_2 + \|\widetilde{\mathbf{E}}\|_{\mathrm{F}} && (9)\\
&\leq \sqrt{\gamma}\|(\mathbf{I}_m - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}})\mathbf{AR}\|_{\mathrm{F}}\|(\mathbf{V}_k\mathbf{R})^{\dagger}\|_2 + \|\widetilde{\mathbf{E}}\|_{\mathrm{F}} && (10)\\
&\leq \sqrt{\gamma}\sqrt{1+\varepsilon}\|(\mathbf{I}_m - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}})\mathbf{A}\|_{\mathrm{F}}\frac{1}{1-\varepsilon} \\
&\quad + 3\varepsilon\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}} && (11)\\
&\leq \sqrt{\gamma}(1+2.5\varepsilon)\|(\mathbf{I}_m - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}})\mathbf{A}\|_{\mathrm{F}} \\
&\quad + 3\varepsilon\sqrt{\gamma}\|(\mathbf{I}_m - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}})\mathbf{A}\|_{\mathrm{F}} && (12)\\
&= \sqrt{\gamma}(1+5.5\varepsilon)\|(\mathbf{I}_m - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}})\mathbf{A}\|_{\mathrm{F}} && (13)
\end{aligned}$$

In Eqn. (8), we used the second statement of Lemma 7, the triangle inequality for matrix norms, and the fact that $\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}$ is a projection matrix and can be dropped without increasing the Frobenius norm. In Eqn. (9), we used spectral submultiplicativity and the fact that $\mathbf{V}_k^{\mathrm{T}}$ can be dropped without changing the spectral norm. In Eqn. (10), we replaced $\mathbf{X}_{\tilde{\gamma}}$ by $\mathbf{X}_{\mathrm{opt}}$ and the factor $\sqrt{\gamma}$ appeared in the first term. To better understand this step, notice that $\mathbf{X}_{\tilde{\gamma}}$ gives a γ-approximation to the optimal k-means clustering of the matrix $\mathbf{C}$, and any other $m \times k$ indicator matrix (for example, the matrix $\mathbf{X}_{\mathrm{opt}}$)

satisfies,

$$\| \left(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\right)\mathbf{C}\|_{\mathrm{F}}^2 \leq \gamma \min_{\mathbf{X}\in\mathcal{X}} \|(\mathbf{I}_m - \mathbf{X}\mathbf{X}^{\mathrm{T}})\mathbf{C}\|_{\mathrm{F}}^2 \leq \gamma \| \left(\mathbf{I}_m - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\right)\mathbf{C}\|_{\mathrm{F}}^2.$$

In Eqn. (11), we used the first statement of Lemma 7 and Lemma 6 with $\mathbf{Y} = (\mathbf{I} - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}})\mathbf{A}$. In Eqn. (12), we used the fact that $\gamma \geq 1$, the optimality of SVD, and that for any $\varepsilon \in (0, 1/3)$, $\sqrt{1+\varepsilon}/(1-\varepsilon) \leq 1 + 2.5\varepsilon$. Taking squares in Eqn. (13) we obtain,

$$\theta_3^2 \leq \gamma\ (1+5.5\varepsilon)^2\ \|(\mathbf{I}_m - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}})\mathbf{A}\|_{\mathrm{F}}^2 \leq \gamma\,(1+15\varepsilon)\|(\mathbf{I}_m - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}})\mathbf{A}\|_{\mathrm{F}}^2.$$

Rescaling ε accordingly gives the approximation bound in the theorem ($c_2 = 3330 \cdot 15^2$). The failure probability $0.04 + \delta_\gamma$ follows by a union bound on the failure probability δ_γ of the γ-approximation k-means algorithm (Definition 2), Lemma 6, and Lemma 7. ■

A. Disscusion

As we mentioned in Section I-A, one can project the data down to $O(\log(m)/\varepsilon^2)$ dimensions and guarantee a clustering error which is not more than $(1+\varepsilon)$ times the optimal clustering error. This result is straightforward using the Johnson-Lindenstrauss lemma, which asserts that after such a dimension reduction all pairwise (Euclidian) distances of the points would be preserved by a factor $(1+\varepsilon)$ [15]. If distances are preserved, then all clusterings - hence the optimal one - are preserved by the same factor.

Our result here extends the Johnson-Lindenstrauss result in a remarkable way. It argues that much less dimensions suffice to preserve the optimal clustering in the data. We do not prove that pairwise distances are preserved. Our proof uses the linear algebraic formulation of the k-means clustering problem and shows that if the spectral information of certain matrices is preserved then the k-means clustering is preserved as well. Our bound is worse than the relative error bound obtained with $O(\log(m)/\varepsilon^2)$ dimensions; we believe though that it is possible to obtain a relative error bound and the $(2+\varepsilon)$ bound might be an artifact of the analysis.

VI. Feature Extraction With Approximate SVD

Finally, we present a feature extraction algorithm that employs the SVD to construct $r = k$ artificial features. Our method and proof techniques are the same with those of [12] with the only difference being the fact that we use a fast approximate (randomized) SVD via Lemma 2 as opposed to the expensive exact deterministic SVD. In fact, replacing $\mathbf{Z} = \mathbf{V}_k$ reproduces the proof in [12]. Our choice gives a considerably faster algorithm with approximation error comparable to the error in [12].

Theorem 3: Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and k be inputs of the k-means clustering problem. Let $\varepsilon \in (0,1)$ and construct features $\mathbf{C} \in \mathbb{R}^{m\times k}$ by using Algorithm 3 in $O(mnk/\varepsilon)$ time. Run any γ-approximation k-means algorithm with failure probability δ_γ on $\mathbf{C}, k$ and construct $\mathbf{X}_{\tilde{\gamma}}$. Then, w.p. at least $0.99 - \delta_\gamma$,

$$\|\mathbf{A} - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2 \leq (1 + (1+\varepsilon)\gamma\,)\,\|\mathbf{A} - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2.$$

Algorithm 3 Randomized Feature Extraction for k-Means Clustering

Input: Dataset $\mathbf{A} \in \mathbb{R}^{m\times n}$, number of clusters k, and $0 < \varepsilon < 1$.
Output: $\mathbf{C} \in \mathbb{R}^{m\times k}$ with k artificial features.

1: Let $\mathbf{Z} = \mathrm{FastFrobeniusSVD}(\mathbf{A}, k, \varepsilon)$; $\mathbf{Z} \in \mathbb{R}^{n\times k}$ (via Lemma 2).
2: Return $\mathbf{C} = \mathbf{A}\mathbf{Z} \in \mathbb{R}^{m\times k}$.

In words, given any set of points in some n-dimensional space and the number of clusters k, it suffices to create exactly k new features (via an approximate Singular Value Decomposition) and then run some k-means algorithm on this new dataset. The theorem formally argues that the clustering it would be obtained in the low-dimensional space will be close to the clustering it would have been obtained after running the k-means method in the original high-dimensional data. We also state the result of the theorem in the notation we introduced in Section I: $\mathcal{F}(\mathcal{P}, \mathcal{S}_{\tilde{\gamma}}) \leq (1 + (1+\varepsilon)\gamma\,)\,\mathcal{F}(\mathcal{P}, \mathcal{S}_{opt})$. Here, $\mathcal{S}_{\tilde{\gamma}}$ is the partition obtained after running the γ-approximation k-means algorithm on the low-dimensional space. The approximation factor is $(1 + (1+\varepsilon)\gamma\,)$. The term $\gamma > 1$ is due to the fact that the k-means method that we run in the low-dimensional space does not recover the optimal k-means partition. The other factor $1+\varepsilon$ is due to the fact that we run k-means in the low-dimensional space.

Proof (of Theorem 3): We start by manipulating the term $\|\mathbf{A} - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2$. Notice that $\mathbf{A} = \mathbf{A}\mathbf{Z}\mathbf{Z}^{\mathrm{T}} + \mathbf{E}$. Also, $\left(\left(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\right)\mathbf{A}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\right)\left(\left(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\right)\mathbf{E}\right)^{\mathrm{T}} = \mathbf{0}_{m\times m}$, because $\mathbf{Z}^{\mathrm{T}}\mathbf{E}^{\mathrm{T}} = \mathbf{0}_{k\times m}$, by construction. Now, using the Matrix Pythagorean theorem (see Lemma 1 in Section III),

$$\|\mathbf{A} - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2 = \underbrace{\|(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}})\mathbf{A}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\|_{\mathrm{F}}^2}_{\theta_1^2} + \underbrace{\|(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}})\mathbf{E}\|_{\mathrm{F}}^2}_{\theta_2^2}. \quad (14)$$

We first bound the second term of Eqn. (14). Since $\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}$ is a projection matrix, it can be dropped without increasing the Frobenius norm (see Section III). Applying Markov's inequality on the equation of Lemma 2, we obtain that w.p. 0.99,

$$\|\mathbf{E}\|_{\mathrm{F}}^2 \leq (1+100\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}^2. \quad (15)$$

(This is Eqn. 2, of which we provided a short proof in the Appendix.) Note also that $\mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{A}$ has rank at most k; so, from the optimality of the SVD, overall,

$$\theta_2^2 \leq (1+100\varepsilon)\|\mathbf{A} - \mathbf{A}_k\|_{\mathrm{F}}^2 \leq (1+100\varepsilon)\|\mathbf{A} - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2 = (1+100\varepsilon)\mathrm{F}_{\mathrm{opt}}.$$

Hence, it follows that w.p. 0.99,

$$\theta_2^2 \leq (1 + 100\varepsilon)\mathrm{F}_{\mathrm{opt}}.$$

We now bound the first term in Eqn. (14),

$$\theta_1 \leq \|(\mathbf{I}_m - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}})\mathbf{A}\mathbf{Z}\|_{\mathrm{F}} \quad (16)$$

$$\leq \sqrt{\gamma}\|(\mathbf{I}_m - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}})\mathbf{A}\mathbf{Z}\|_{\mathrm{F}} \quad (17)$$

$$\leq \sqrt{\gamma}\|(\mathbf{I}_m - \mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{\mathrm{T}})\mathbf{A}\|_{\mathrm{F}} \quad (18)$$

In Eqn. (16), we used spectral submultiplicativity and the fact that $\|\mathbf{Z}^{\mathrm{T}}\|_2 = 1$. In Eqn. (17), we replaced $\mathbf{X}_{\tilde{\gamma}}$ by $\mathbf{X}_{\mathrm{opt}}$ and the factor $\sqrt{\gamma}$ appeared in the first term (similar argument as in the proof of Theorem 1). In Eqn. (18), we used spectral submultiplicativity and the fact that $\|\mathbf{Z}\|_2 = 1$. Overall (assuming $\gamma \geq 1$),

$$\|\mathbf{A} - \mathbf{X}_{\tilde{\gamma}}\mathbf{X}_{\tilde{\gamma}}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2 \leq \theta_1^2 + \theta_2^2 \leq \gamma\,\mathrm{F}_{\mathrm{opt}} + (1 + 100\varepsilon)\mathrm{F}_{\mathrm{opt}}$$
$$\leq \mathrm{F}_{\mathrm{opt}} + (1 + 10^2\varepsilon)\gamma\,\mathrm{F}_{\mathrm{opt}}.$$

The failure probability is $0.01 + \delta_\gamma$, from a union bound on Lemma 2 and Definition 2. Finally, rescaling ε accordingly gives the approximation bound in the theorem. ■

VII. Experiments

This section describes a preliminary experimental evaluation of the feature selection and feature extraction algorithms presented in this work. We implemented the proposed algorithms in MATLAB [26] and compared them against a few other prominent dimensionality reduction techniques such as the Laplacian scores [27]. Laplacian scores is a popular feature selection method for clustering and classification. We performed all the experiments on a Mac machine with a dual core 2.8 Ghz processor and 8 GB of RAM.

Our empirical findings are far from exhaustive, however they indicate that the feature selection and feature extraction algorithms presented in this work achieve a satisfactory empirical performance with rather small values of r (far smaller than the theoretical bounds presented here). We believe that the large constants that appear in our theorems (see Theorem 2) are artifacts of our theoretical analysis and can be certainly improved.

A. Dimensionality Reduction Methods

Given m points described with respect to n features and the number of clusters k, our goal is to select or construct r features on which we execute Lloyd's algorithm for k-means on this constructed set of features. In this section, we experiment with various methods for selecting or constructing the features. The number of features to be selected or extracted is part of the input as well. In particular, in Algorithm 1 we do not consider ε to be part of the input. We test the performance of the proposed algorithms for various values of r, and we compare our algorithms against other feature selection and feature extraction methods from the literature, that we summarize below:

1) **Randomized Sampling with Exact SVD (Sampl/SVD).** This corresponds to Algorithm 1 with the following modification. In the first step of the algorithm, the matrix $\mathbf{Z}$ is calculated to contain exactly the top k right singular vectors of $\mathbf{A}$.
2) **Randomized Sampling with Approximate SVD (Sampl/ApproxSVD).** This corresponds to Algorithm 1 with ε fixed to $1/3$.
3) **Random Projections (RP).** Here we use Algorithm 2. However, in our implementation we use the naive approach for the matrix-matrix multiplication in the third step (not the Mailman algorithm [25]).
4) **SVD.** This is Algorithm 3 with the following modification. In the first step of the algorithm, the matrix $\mathbf{Z}$ is calculated to contain exactly the top k right singular vectors of $\mathbf{A}$.
5) **Approximate SVD (ApprSVD).** This corresponds to Algorithm 3 with ε fixed to $1/3$.
6) **Laplacian Scores (LapScores).** This corresponds to the feature selection method described in [27]. We use the MATLAB code from [28] with the default parameters. In particular, in MATLAB notation we executed the following commands,

$$\mathbf{W} = \mathrm{constructW}(\mathbf{A});\ \mathrm{Scores} = \mathrm{LaplacianScore}(\mathbf{A}, \mathbf{W});$$

Finally, we also compare all these methods against evaluating the k-means algorithm in the full dimensional dataset which we denote by `kMeans`.

B. k-Means Method

Although our theorems allow the use of any γ-approximation algorithm for k-means, in practice the Lloyd's algorithm performs very well [2]. Hence, we employ the Lloyd's algorithm in our experiments. Namely, every time we mention *"we run k-means"*, we mean that we run 500 iterations of the Lloyd's algorithm with 5 different random initializations and return the best outcome over all repetitions, i.e., in MATLAB notation we run the following command, `kmeans(A, k, 'Replicates', 5, 'Maxiter', 500)`.

C. Datasets

We performed experiments on a few real-world and synthetic datasets. For the synthetic dataset, we generated a dataset of $m = 1000$ points in $n = 2000$ dimensions as follows. We chose $k = 5$ centers uniformly at random from the n-dimensional hypercube of side length 2000 as the ground truth centers. We then generated points from a Gaussian distribution of variance one, centered at each of the real centers. To each of the 5 centers we generated 200 points (we did not include the centers in the dataset). Thus, we obtain a number of well separated Gaussians with the real centers providing a good approximation to the optimal clustering. We will refer to this dataset as `Synth`.

For the real-world datasets we used five datasets that we denote by `USPS`, `COIL20`, `ORL`, `PIE` and `LIGHT`. The `USPS` digit dataset contains grayscale pictures of handwritten digits and can be downloaded from the UCI repository [29]. Each data point of `USPS` has 256 dimensions and there are

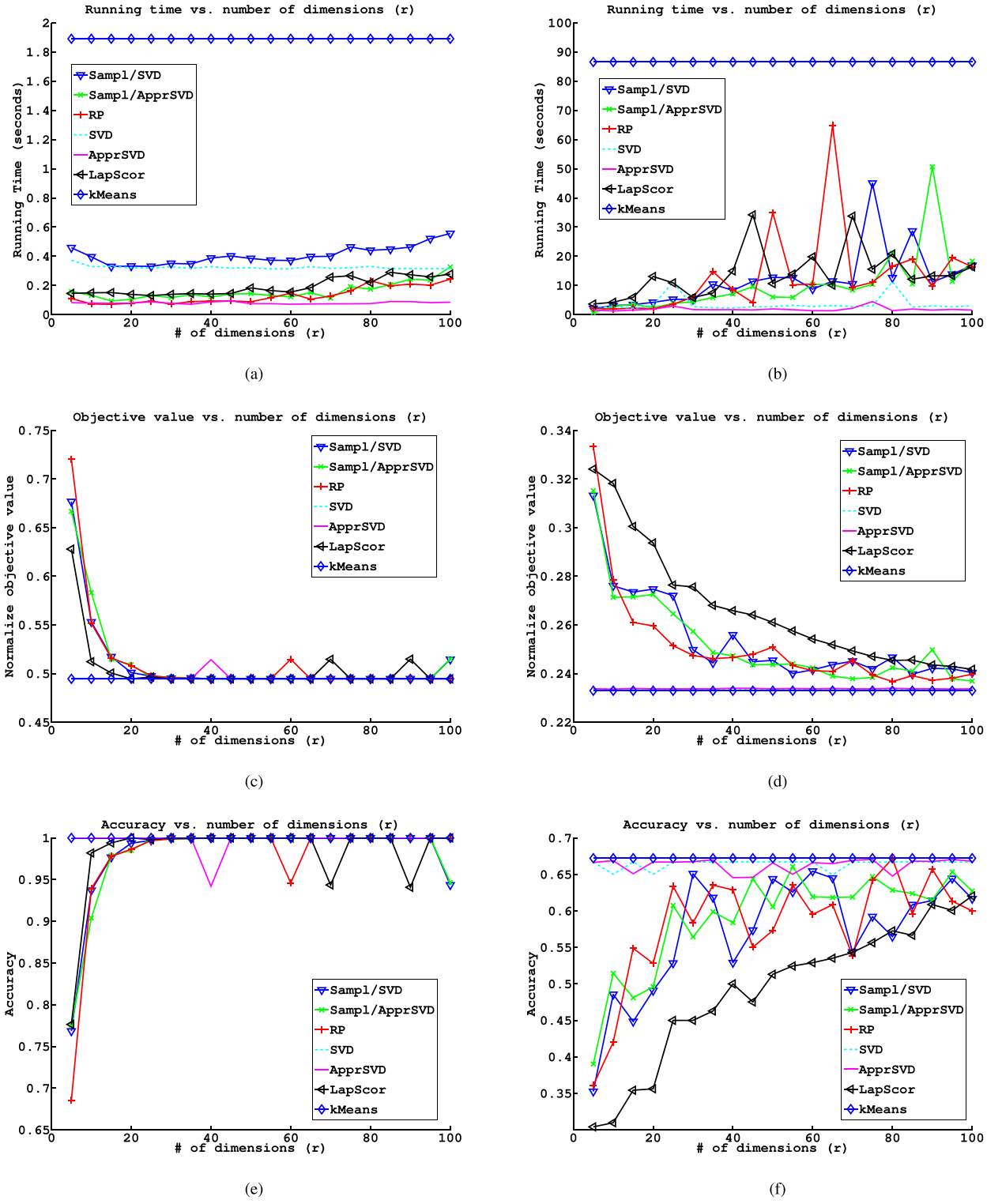

Fig. 1. Plot of running time (a), (b), objective value (c), (d) and accuracy (e), (f) versus the number of projected dimensions for several dimensionality reduction approaches. Left column corresponds to the Synth dataset, whereas the right column corresponds to the USPS dataset.

1100 data points per digit. The coefficients of the data points have been normalized between 0 and 1. The COIL20 dataset contains 1400 images of 20 objects (the images of each objects were taken 5 degrees apart as the object is rotated on a turntable and each object has 72 images) and can be downloaded from [30]. The size of each image is 32×32 pixels, with 256 grey levels per pixel. Thus, each image is represented by a 1024-dimensional vector. ORL contains ten different images each of 40 distinct subjects and can be located at [31]. For few subjects, the images were taken at different times, varying the lighting, facial expressions and facial details. All the images were taken against

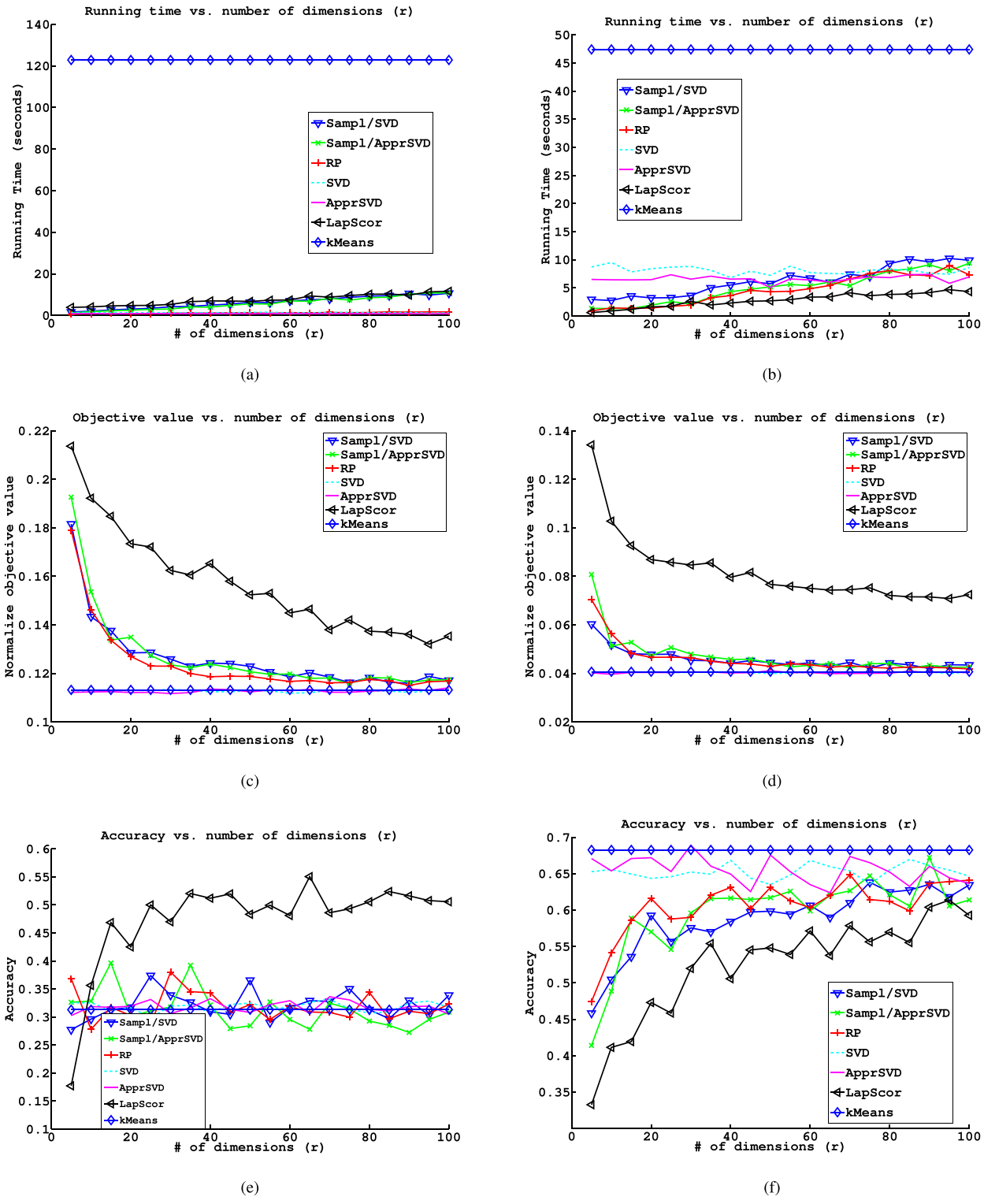

Fig. 2. Plot of running time (a), (b), objective value (c), (d) and accuracy (e), (f) versus the number of projected dimensions for several dimensionality reduction approaches. Left column corresponds to the COIL20 dataset, whereas the right column corresponds to the LIGHT dataset.

a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). There are in total 400 different objects having 4096 dimensions.

PIE is a database of 41,368 images of 68 people, each person under 13 different poses, 43 different illumination conditions, and with 4 different expressions [32]. Our dataset contains only five near frontal poses (C05, C07, C09, C27, C29) and all the images under different illuminations and expressions. Namely, there are in total 2856 data points with 1024 dimensions. The LIGHT dataset is identical with the dataset that has been used in [27], the data points of

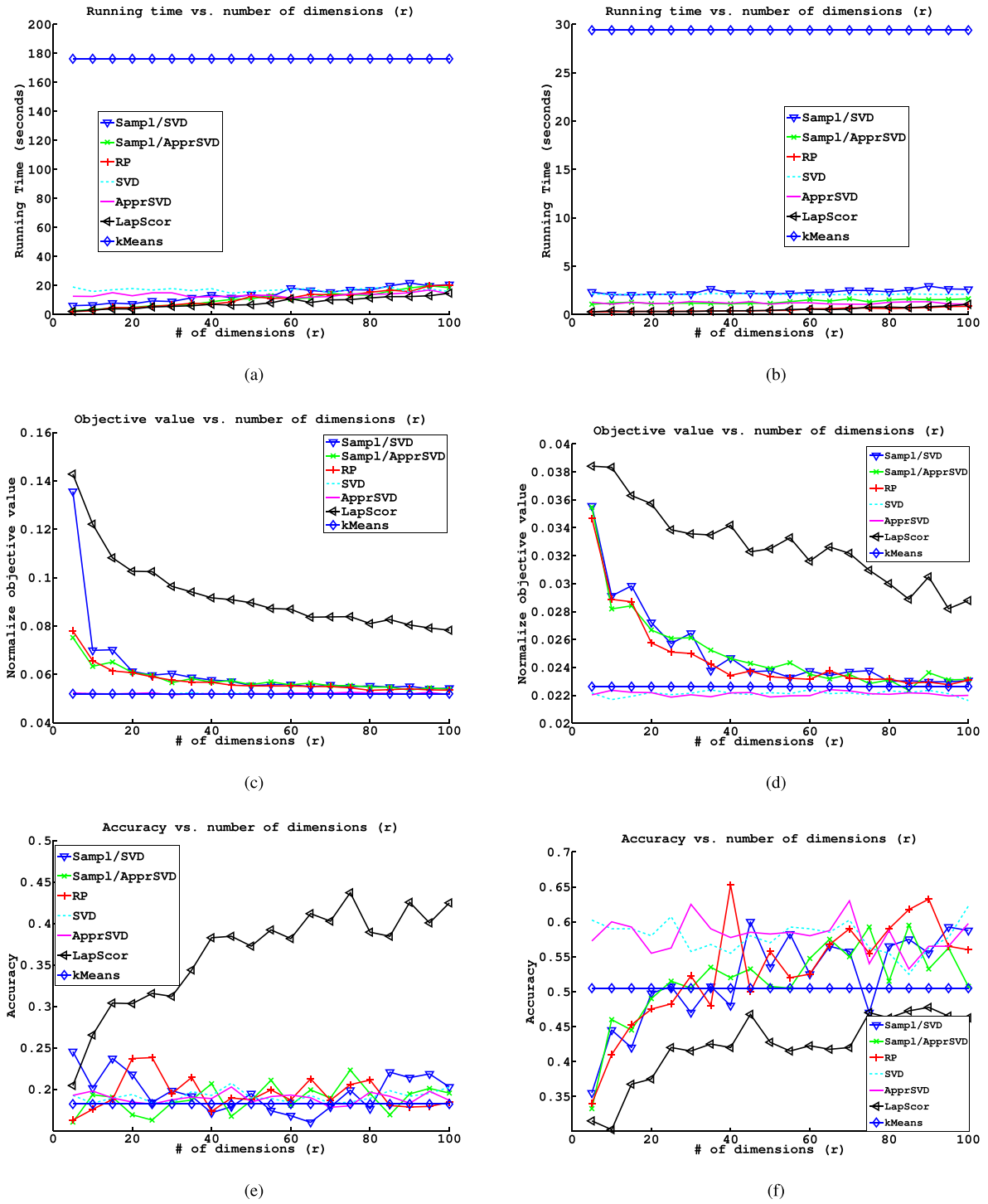

Fig. 3. Plot of running time (a), (b), objective value (c), (d) and accuracy versus (e), (f) the number of projected dimensions for several dimensionality reduction approaches. Left column corresponds to the PIE dataset, whereas the right column corresponds to the ORL dataset.

LIGHT is 1428 containing 1014 features. For each real-world dataset we fixed k to be equal to the cardinality of their corresponding label set.

D. Evaluation Methodology

As a measure of quality of all methods we measure and report the objective function $\mathcal{F}$ of the k-means clustering problem. In particular, we report a normalized version of $\mathcal{F}$, i.e. $\mathcal{F} = \mathcal{F}/\|\mathbf{A}\|_{\mathrm{F}}^2$. In addition, we report the misclassification accuracy of the clustering result based on the labelled information of the input data. We denote this number by P $(0 \leq P \leq 1)$, where $P = 0.9$, for example, implies that 90% of the points were assigned to the "correct cluster"/label after the application of the clustering algorithm. Finally, we report running times (in seconds). It is important to highlight that we report the running time of both the dimensionality reduction procedure and the k-means algorithm applied on the

low-dimensional projected space for all proposed algorithms. All the reported quantities correspond to the average values of five independent executions.

E. Results

We present the results of our experiments in Figs. 1–3. We experimented with relative small values for the number of dimensions:

$$r = 5, 10, 15, \ldots, 100.$$

In the synthetic dataset, we observe that all dimensionality reduction methods for k-means clustering are clearly more efficient compared to naive k-means clustering. More importantly, the accuracy plots of Figure 1 demonstrate that the dimensionality reduction approach is also accurate in this case even for relatively (with respect to k) small values of r, i.e., $\approx$ 20. Recall that in this case the clusters of the dataset are well-separated between each other. Hence, these observations suggest that dimensionality reduction for k-means clustering is effective when applied to well-separated data points.

The behavior of the dimensionality reduction methods for k-means clustering for the real-world datasets is similar with the synthetic dataset, see Figures 2 and 3. That is, as the number of projecting dimensions increases, the normalized objective value of the resulting clustering decreases. Moreover, in all cases the normalized objective value of the proposed methods converge to the objective value attained by the naive k-means algorithm (as the number of dimensions increases). In all cases but the `PIE` and `COIL20` dataset, the proposed dimensionality reduction methods have superior performance compared to Laplacian Scores [27] both in terms of accuracy and normalized k-means objective value. In the `PIE` and `COIL20` datasets, the Laplacian Scores approach is superior compared to all other approaches in terms of accuracy. However, notice that in these two datasets the naive k-means algorimhm performs poorly in terms of accuracy which indicates that the data might not be well-separated.

Regarding the running times of the algorithms notice that in some cases the running time does not necessarily increased by increasing the number of dimensions. This happens because after the dimensionality reduction step the k-means method might take a different number of iterations to converge. We did not investigated this behavior further since this is not the focus of our experimental evaluation.

Our experiments indicate that running our algorithms with small values of r, e.g., $r = 20$ or $r = 30$, achieves nearly optimal separation of a mixture of Gaussians and does well in several real-world clustering problems. Although a more thorough experimental evaluation of our algorithms would have been far more informative, our preliminary experimental findings are quite encouraging with respect to the performance of our algorithms in *practice*.

VIII. Conclusions

We studied the problem of dimensionality reduction for k-means clustering. Most of the existing results in this topic consist of heuristic approaches, whose excellent empirical performance can not be explained with a rigorous theoretical analysis. In this paper, our focus was on dimensionality reduction methods that work well *in theory*. We presented three such approaches, one feature selection method for k-means and two feature extraction methods. The theoretical analysis of the proposed methods is based on the fact that dimensionality reduction for k-means has deep connections with low-rank approximations to the data matrix that contains the points one wants to cluster. We explained those connections in the text and employed modern fast algorithms to compute such low rank approximations and designed fast algorithms for dimensionality reduction in k-means.

Despite our focus on the theoretical foundations of the proposed algorithms, we tested the proposed methods in practice and concluded that the experimental results are very encouraging: dimensionality reduction for k-means using the proposed techniques leads to faster algorithms that are almost as accurate as running k-means on the high dimensional data.

All in all, our work describes the *first* provably efficient feature selection algorithm for k-means clustering as well as two novel provably efficient feature extraction algorithms. An interesting path for future research is to design provably efficient $(1 + \varepsilon)$-error dimensionality reduction methods for k-means.

Appendix
Technical Lemmata

The following technical lemma is useful in the proof of Lemma 5 and the proof of Lemma 7.

Lemma 8: Let $\mathbf{Q} \in \mathbb{R}^{n\times k}$ *with* $n > k$ *and* $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}_k$. *Let* Θ *be any* $n \times r$ *matrix* $(r > k)$ *satisfying* $1-\varepsilon \leq \sigma_i^2(\mathbf{Q}^{\mathrm{T}}\Theta) \leq 1+\varepsilon$ *for every* $i = 1, \ldots, k$ *and* $0 < \varepsilon < 1/3$. *Then,*

$$\|(\mathbf{Q}^{\mathrm{T}}\Theta)^{\dagger} - (\mathbf{Q}^{\mathrm{T}}\Theta)^{\mathrm{T}}\|_2 \leq \frac{\varepsilon}{\sqrt{1-\varepsilon}} \leq 1.5\varepsilon.$$

Proof: Let $\mathbf{X} = \mathbf{Q}^{\mathrm{T}}\Theta \in \mathbb{R}^{k\times r}$ with SVD $\mathbf{X} = \mathbf{U}_{\mathbf{X}}\Sigma_{\mathbf{X}}\mathbf{V}_{\mathbf{X}}^{\mathrm{T}}$. Here, $\mathbf{U}_{\mathbf{X}} \in \mathbb{R}^{k\times k}$, $\Sigma_{\mathbf{X}} \in \mathbb{R}^{k\times k}$, and $\mathbf{V}_{\mathbf{X}} \in \mathbb{R}^{r\times k}$, since $r > k$. Consider taking the SVD of $(\mathbf{Q}^{\mathrm{T}}\Theta)^{\dagger}$ and $(\mathbf{Q}^{\mathrm{T}}\Theta)^{\mathrm{T}}$,

$$\begin{aligned}\|(\mathbf{Q}^{\mathrm{T}}\Theta)^{\dagger} - (\mathbf{Q}^{\mathrm{T}}\Theta)^{\mathrm{T}}\|_2 &= \|\mathbf{V}_{\mathbf{X}}\Sigma_{\mathbf{X}}^{-1}\mathbf{U}_{\mathbf{X}}^{\mathrm{T}} - \mathbf{V}_{\mathbf{X}}\Sigma_{\mathbf{X}}\mathbf{U}_{\mathbf{X}}^{\mathrm{T}}\|_2 \\ &= \|\mathbf{V}_{\mathbf{X}}(\Sigma_{\mathbf{X}}^{-1} - \Sigma_{\mathbf{X}})\mathbf{U}_{\mathbf{X}}^{\mathrm{T}}\|_2 \\ &= \|\Sigma_{\mathbf{X}}^{-1} - \Sigma_{\mathbf{X}}\|_2,\end{aligned}$$

since $\mathbf{V}_{\mathbf{X}}$ and $\mathbf{U}_{\mathbf{X}}^{\mathrm{T}}$ can be dropped without changing the spectral norm. Let $\mathbf{Y} = \Sigma_{\mathbf{X}}^{-1} - \Sigma_{\mathbf{X}} \in \mathbb{R}^{k\times k}$ be a diagonal matrix. Then, for all $i = 1, \ldots, k$, $\mathbf{Y}_{ii} = \frac{1-\sigma_i^2(\mathbf{X})}{\sigma_i(\mathbf{X})}$. Since $\mathbf{Y}$ is diagonal,

$$\begin{aligned}\|\mathbf{Y}\|_2 &= \max_{1\leq i\leq k} |\mathbf{Y}_{ii}| \\ &= \max_{1\leq i\leq k} \left|\frac{1-\sigma_i^2(\mathbf{X})}{\sigma_i(\mathbf{X})}\right| \\ &= \max_{1\leq i\leq k} \frac{\left|1-\sigma_i^2(\mathbf{X})\right|}{\sigma_i(\mathbf{X})} \\ &\leq \frac{\varepsilon}{\sqrt{1-\varepsilon}} \leq 1.5\varepsilon.\end{aligned}$$

The first equality follows since the singular values are positive (from our choice of ε and the left hand side of the bound for the singular values). The first inequality follows by the bound for the singular values of $\mathbf{X}$. The last inequality follows by the assumption that $0 < \varepsilon < 1/3$. ∎

Proof of Lemma 5: We begin with the analysis of a matrix-multiplication-type term involving the multiplication of the matrices $\mathbf{E}$, $\mathbf{Z}$. The sampling and rescaling matrices $\mathbf{\Omega}, \mathbf{S}$ indicate the subsampling of the columns and rows of $\mathbf{E}$, $\mathbf{Z}$, respectively. [33, Lemma 4, eq. (4)] gives a bound for such $\mathbf{\Omega}, \mathbf{S}$ constructed with randomized sampling with replacement and *any* set of probabilities $p_1, p_2, \ldots, p_n$ (over the columns of $\mathbf{E}$ - rows of $\mathbf{Z}$),

$$\mathbb{E}\,\|\mathbf{EZ} - \mathbf{E\Omega SS}^{\mathrm{T}}\mathbf{\Omega}^{\mathrm{T}}\mathbf{Z}\|_{\mathrm{F}}^2 \leq \sum_{i=1}^{n} \frac{\|\mathbf{E}^{(i)}\|_2^2 \|\mathbf{Z}_{(i)}\|_2^2}{r p_i} - \frac{1}{r}\|\mathbf{EZ}\|_{\mathrm{F}}^2.$$

Notice that $\mathbf{EZ} = \mathbf{0}_{m\times k}$, by construction (see Lemma 2). Now, for every $i = 1, \ldots, n$ replace the values $p_i = \frac{\|\mathbf{Z}_{(i)}\|_2^2}{k}$ (in Definition 3) and rearrange,

$$\mathbb{E}\,\|\mathbf{E\Omega SS}^{\mathrm{T}}\mathbf{\Omega}^{\mathrm{T}}\mathbf{Z}\|_{\mathrm{F}}^2 \leq \frac{k}{r}\|\mathbf{E}\|_{\mathrm{F}}^2. \tag{19}$$

Observe that Lemma 3 and our choice of r, implies that w.p. $1-\delta$,

$$1 - \varepsilon \leq \sigma_i^2(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S}) \leq 1 + \varepsilon, \quad \text{for all } i = 1, \ldots, k. \tag{20}$$

For what follows, condition on the event of Ineq. (20). First, $\sigma_k(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S}) > 0$. So, $\mathrm{rank}(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S}) = k$ and $(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\dagger} = \mathbf{I}_k$.[3] Now, $\mathbf{AZZ}^{\mathrm{T}} - \mathbf{AZZ}^{\mathrm{T}}\mathbf{\Omega S}(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\dagger}\mathbf{Z}^{\mathrm{T}} = \mathbf{AZZ}^{\mathrm{T}} - \mathbf{AZI}_k\mathbf{Z}^{\mathrm{T}} = \mathbf{0}_{m\times n}$. Next, we manipulate the term $\theta = \|\mathbf{AZZ}^{\mathrm{T}} - \mathbf{A\Omega S}(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\dagger}\mathbf{Z}^{\mathrm{T}}\|_{\mathrm{F}}$ as follows (recall, $\mathbf{A} = \mathbf{AZZ}^{\mathrm{T}} + \mathbf{E}$),

$$\begin{aligned}\theta &= \|\underbrace{\mathbf{AZZ}^{\mathrm{T}} - \mathbf{AZZ}^{\mathrm{T}}\mathbf{\Omega S}(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\dagger}\mathbf{Z}^{\mathrm{T}}}_{\mathbf{0}_{m\times n}} - \mathbf{E\Omega S}(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\dagger}\mathbf{Z}^{\mathrm{T}}\|_{\mathrm{F}} \\ &= \|\mathbf{E\Omega S}(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\dagger}\mathbf{Z}^{\mathrm{T}}\|_{\mathrm{F}}.\end{aligned}$$

Finally, we manipulate the latter term as follows,

$$\begin{aligned}&\|\mathbf{E\Omega S}(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\dagger}\mathbf{Z}^{\mathrm{T}}\|_{\mathrm{F}} \leq \|\mathbf{E\Omega S}(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\dagger}\|_{\mathrm{F}} \\ &\leq \|\mathbf{E\Omega S}(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\mathrm{T}}\|_{\mathrm{F}} + \|\mathbf{E\Omega S}\|_{\mathrm{F}}\|(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\dagger} - (\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\mathrm{T}}\|_2 \\ &\leq \sqrt{\frac{k}{\delta r}}\|\mathbf{E}\|_{\mathrm{F}} + \frac{1}{\sqrt{\delta}}\|\mathbf{E}\|_{\mathrm{F}}\frac{\varepsilon}{\sqrt{1-\varepsilon}} \\ &\leq \left(\sqrt{\frac{k}{\delta r}} + \frac{\varepsilon}{\sqrt{\delta}\sqrt{1-\varepsilon}}\right)\|\mathbf{E}\|_{\mathrm{F}} \\ &\leq \left(\frac{\varepsilon}{2\sqrt{\delta}}\frac{1}{\sqrt{\ln(2k/\delta)}} + \frac{\varepsilon}{\sqrt{\delta}\sqrt{1-\varepsilon}}\right)\|\mathbf{E}\|_{\mathrm{F}} \\ &\leq \left(\frac{\varepsilon}{2\ln(4)\sqrt{\delta}} + \frac{\varepsilon}{\sqrt{\delta}\sqrt{1-\varepsilon}}\right)\|\mathbf{E}\|_{\mathrm{F}} \leq \frac{1.6\varepsilon}{\sqrt{\delta}}\|\mathbf{E}\|_{\mathrm{F}}.\end{aligned}$$

The first inequality follows by spectral submultiplicativity and the fact that $\|\mathbf{Z}^{\mathrm{T}}\|_2 = 1$. The second inequality follows by the triangle inequality for matrix norms. In the third inequality, the bound for the term $\|\mathbf{E\Omega S}(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\mathrm{T}}\|_{\mathrm{F}}$ follows by applying to it Markov's inequality together with Ineq. (19); also, $\|\mathbf{E\Omega S}\|_{\mathrm{F}}$ is bounded by $(1/\sqrt{\delta})\|\mathbf{E}\|_{\mathrm{F}}$ w.p. $1-\delta$ (Lemma 4), while we bound $\|(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\dagger} - (\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\mathrm{T}}\|_2$ using Lemma 8 (set $\mathbf{Q} = \mathbf{Z}$ and $\Theta = \mathbf{\Omega S}$). So, by the union bound, the failure probability is 3δ. The rest of the argument follows by our choice of r, assuming $k \geq 2$, $\varepsilon < 1/3$ and simple algebraic manipulations. ∎

Proof of Lemma 6: First, define the random variable $Y = \|\mathbf{YR}\|_{\mathrm{F}}^2$. It is easy to see that $\mathbb{E}\,Y = \|\mathbf{Y}\|_{\mathrm{F}}^2$ and moreover an upper bound for the variance of Y is available in [18, Lemma 8]: $\mathrm{Var}[Y] \leq 2\|\mathbf{Y}\|_{\mathrm{F}}^4/r$.[4] Now, Chebyshev's inequality tells us that,

$$\begin{aligned}\mathbb{P}(|Y - \mathbb{E}\,Y| \geq \varepsilon\|\mathbf{Y}\|_{\mathrm{F}}^2) &\leq \frac{\mathrm{Var}[Y]}{\varepsilon^2\|\mathbf{Y}\|_{\mathrm{F}}^4} \\ &\leq \frac{2\|\mathbf{Y}\|_{\mathrm{F}}^4}{r\varepsilon^2\|\mathbf{Y}\|_{\mathrm{F}}^4} \\ &\leq \frac{2}{c_0 k} \\ &\leq 0.01.\end{aligned}$$

The last inequality follows by assuming $c_0 \geq 100$ and the fact that $k > 1$. Finally, taking square root on both sides concludes the proof. ∎

Proof of Lemma 7: We start with the definition of the Johnson-Lindenstrauss transform.

Definition 4: (Johnson-Lindenstrauss Transform): A random matrix $\mathbf{R} \in \mathbb{R}^{n\times r}$ forms a Johnson-Lindenstrauss transform if, for any (row) vector $\boldsymbol{x} \in \mathbb{R}^n$,

$$\mathbb{P}\Big((1-\varepsilon)\,\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{x}\mathbf{R}\|_2^2 \leq (1+\varepsilon)\,\|\boldsymbol{x}\|_2^2\Big) \geq 1 - e^{-C\varepsilon^2 r}$$

where $C > 0$ is an absolute constant.

Notice that in order to achieve failure probability at most δ, it suffices to take $r = O(\log(1/\delta)/\varepsilon^2)$. We continue with [24, Th. 1.1] (properly stated to fit our notation and after minor algebraic manipulations), which indicates that a (rescaled) sign matrix $\mathbf{R}$ corresponds to a *Johnson-Lindenstrauss transform* as defined above.

Theorem 4 ([24]):[5] Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $0 < \varepsilon < 1$. Let $\mathbf{R} \in \mathbb{R}^{n\times r}$ be a rescaled random sign matrix with $r = \frac{36}{\varepsilon^2}\log(m)\log(1/\delta)$. Then for all $i, j = 1, \ldots, m$ and w.p. at least $1-\delta$,

$$\begin{aligned}(1-\varepsilon)\|\mathbf{A}_{(i)} - \mathbf{A}_{(j)}\|_2^2 &\leq \|\left(\mathbf{A}_{(i)} - \mathbf{A}_{(j)}\right)\mathbf{R}\|_2^2 \\ &\leq (1+\varepsilon)\|\mathbf{A}_{(i)} - \mathbf{A}_{(j)}\|_2^2.\end{aligned}$$

In addition, we will use a matrix multiplication bound which follows from [18, Lemma 6]. The second claim of this lemma says that for any $\mathbf{X} \in \mathbb{R}^{m\times n}$ and $\mathbf{Y} \in \mathbb{R}^{n\times p}$, if $\mathbf{R} \in \mathbb{R}^{n\times r}$

[3] To see this, let $\mathbf{B} = \mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S} \in \mathbb{R}^{k\times r}$ with SVD $\mathbf{B} = \mathbf{U_B}\mathbf{\Sigma_B}\mathbf{V_B}^{\mathrm{T}}$. Here, $\mathbf{U_B} \in \mathbb{R}^{k\times k}$, $\mathbf{\Sigma_B} \in \mathbb{R}^{k\times k}$, and $\mathbf{V_B} \in \mathbb{R}^{r\times k}$, since $r > k$. Finally, $(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})(\mathbf{Z}^{\mathrm{T}}\mathbf{\Omega S})^{\dagger} = \mathbf{U_B}\mathbf{\Sigma_B}\underbrace{\mathbf{V_B}^{\mathrm{T}}\mathbf{V_B}}_{\mathbf{I}_k}\mathbf{\Sigma_B}^{-1}\mathbf{U_B}^{\mathrm{T}} = \mathbf{U_B}\underbrace{\mathbf{\Sigma_B}\mathbf{\Sigma_B}^{-1}}_{\mathbf{I}_k}\mathbf{U_B}^{\mathrm{T}} = \mathbf{I}_k$.

[4] [18] assumes that the matrix $\mathbf{R}$ has i.i.d rows, each one containing four-wise independent zero-mean $\{1/\sqrt{r}, -1/\sqrt{r}\}$ entries. The claim in our lemma follows because our rescaled sign matrix $\mathbf{R}$ satisfies the four-wise independence assumption, by construction.

[5] This theorem is proved by first showing that a rescaled random sign matrix is a *Johnson-Lindenstrauss transform* [24, Lemma 5.1] with constant $C = 36$. Then, setting an appropriate value for r and applying the union bound over all pairs of row indices of $\mathbf{A}$ concludes the proof.

is a matrix with i.i.d rows, each one containing four-wise independent zero-mean $\{1/\sqrt{r}, -1/\sqrt{r}\}$ entries, then,

$$\mathbb{E}\,\|\mathbf{XY} - \mathbf{XRR}^{\mathrm{T}}\mathbf{Y}\|_{\mathrm{F}}^2 \leq \frac{2}{r}\|\mathbf{X}\|_{\mathrm{F}}^2\|\mathbf{Y}\|_{\mathrm{F}}^2. \tag{21}$$

Our random matrix $\mathbf{R}$ uses full independence, hence the above bound holds by dropping the limited independence condition.

1) Statement 1: The first statement in our lemma has been proved in [18, Corollary 11], see also [34, Th. 1.3] for a restatement. More precisely, repeat the proof of [18, Corollary 11] paying attention to the constants. That is, set $\mathbf{C} = \mathbf{V}_k^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{R}\mathbf{V}_k - \mathbf{I}_k$ and $\varepsilon_0 = 1/2$ in [18, Lemma 10], and apply our JL transform with (rescaled) accuracy $\varepsilon/4$ on each vector of the set $T' := \{\mathbf{V}_k^{\mathrm{T}}\boldsymbol{x} \mid \boldsymbol{x} \in T\}$ (which is of size at most $\leq e^{k\ln(18)}$, see [35, Lemma 4] for this bound). So,

$$\mathbb{P}\left(\forall i = 1,\ldots,k:\ 1-\varepsilon \leq \sigma_i^2(\mathbf{V}_k^{\mathrm{T}}\mathbf{R}) \leq 1+\varepsilon\right) \geq 1 - e^{k\ln(18)}e^{-\varepsilon^2 r/(36\cdot 16)}. \tag{22}$$

Setting r such that the failure probability is atmost 0.01 indicates that r should be at least $r \geq 576(k\ln(18)+\ln(100))/\varepsilon^2$. So, $c_0 = 3330$ is a sufficiently large constant for the lemma.

2) Statement 2: Consider the following three events (w.r.t. the randomness of the random matrix $\mathbf{R}$): $\mathcal{E}_1 := \{1-\varepsilon \leq \sigma_i^2(\mathbf{V}_k^{\mathrm{T}}\mathbf{R}) \leq 1+\varepsilon\}$, $\mathcal{E}_2 := \{\|\mathbf{A}_{\rho-k}\mathbf{R}\|_{\mathrm{F}}^2 \leq (1+\varepsilon)\|\mathbf{A}_{\rho-k}\|_{\mathrm{F}}^2\}$ and $\mathcal{E}_3 := \{\|\mathbf{A}_{\rho-k}\mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{V}_k\|_{\mathrm{F}}^2 \leq \varepsilon^2\|\mathbf{A}_{\rho-k}\|_{\mathrm{F}}^2\}$. Ineq. (22) and Lemma 6 with $\mathbf{Y} = \mathbf{A}_{\rho-k}$ imply that $\mathbb{P}(\mathcal{E}_1) \geq 0.99$, $\mathbb{P}(\mathcal{E}_2) \geq 0.99$, respectively. A crucial observation for bounding the failure probability of the last event $\mathcal{E}_3$ is that $\mathbf{A}_{\rho-k}\mathbf{V}_k = \mathbf{U}_{\rho-k}\boldsymbol{\Sigma}_{\rho-k}\mathbf{V}_{\rho-k}^{\top}\mathbf{V}_k = \mathbf{0}_{m\times k}$ by orthogonality of the columns of $\mathbf{V}_k$ and $\mathbf{V}_{\rho-k}$. This event can now be bounded by applying Markov's Inequality on Ineq. (21) with $\mathbf{X} = \mathbf{A}_{\rho-k}$ and $\mathbf{Y} = \mathbf{V}_k$ and recalling that $\|\mathbf{V}_k\|_{\mathrm{F}}^2 = k$ and $r = c_0 k/\varepsilon^2$. Assuming $c_0 \geq 200$, it follows that $\mathbb{P}(\mathcal{E}_3) \geq 0.99$ (hence, setting $c_0 = 3330$ is a sufficiently large constant for both statements). A union bound implies that these three events happen w.p. 0.97. For what follows, condition on these three events.

Let $\widetilde{\mathbf{E}} = \mathbf{A}_k - (\mathbf{AR})(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\dagger}\mathbf{V}_k^{\mathrm{T}} \in \mathbb{R}^{m\times n}$. By setting $\mathbf{A} = \mathbf{A}_k + \mathbf{A}_{\rho-k}$ and using the triangle inequality,

$$\|\widetilde{\mathbf{E}}\|_{\mathrm{F}} \leq \|\mathbf{A}_k - \mathbf{A}_k\mathbf{R}(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\dagger}\mathbf{V}_k^{\mathrm{T}}\|_{\mathrm{F}} + \|\mathbf{A}_{\rho-k}\mathbf{R}(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\dagger}\mathbf{V}_k^{\mathrm{T}}\|_{\mathrm{F}}.$$

The event $\mathcal{E}_1$ implies that $\mathrm{rank}(\mathbf{V}_k^{\mathrm{T}}\mathbf{R}) = k$ thus,[6]

$$(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\dagger} = \mathbf{I}_k.$$

Replacing $\mathbf{A}_k = \mathbf{U}_k\boldsymbol{\Sigma}_k\mathbf{V}_k^{\mathrm{T}}$ and setting $(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\dagger} = \mathbf{I}_k$, we obtain that

$$\begin{aligned}\|\mathbf{A}_k - \mathbf{A}_k\mathbf{R}(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\dagger}\mathbf{V}_k^{\mathrm{T}}\|_{\mathrm{F}} &= \|\mathbf{A}_k - \mathbf{U}_k\boldsymbol{\Sigma}_k\underbrace{\mathbf{V}_k^{\mathrm{T}}\mathbf{R}(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\dagger}}_{\mathbf{I}_k}\mathbf{V}_k^{\mathrm{T}}\|_{\mathrm{F}}\\ &= \|\mathbf{A}_k - \mathbf{U}_k\boldsymbol{\Sigma}_k\mathbf{V}_k^{\mathrm{T}}\|_{\mathrm{F}} = 0.\end{aligned}$$

[6]To see this, let $\mathbf{B} = \mathbf{V}_k^{\mathrm{T}}\mathbf{R} \in \mathbb{R}^{k\times r}$ with SVD $\mathbf{B} = \mathbf{U}_{\mathbf{B}}\boldsymbol{\Sigma}_{\mathbf{B}}\mathbf{V}_{\mathbf{B}}^{\mathrm{T}}$. Here, $\mathbf{U}_{\mathbf{B}} \in \mathbb{R}^{k\times k}$, $\boldsymbol{\Sigma}_{\mathbf{B}} \in \mathbb{R}^{k\times k}$, and $\mathbf{V}_{\mathbf{B}} \in \mathbb{R}^{r\times k}$, since $r > k$. Finally, $(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\dagger} = \mathbf{U}_{\mathbf{B}}\boldsymbol{\Sigma}_{\mathbf{B}}\underbrace{\mathbf{V}_{\mathbf{B}}^{\mathrm{T}}\mathbf{V}_{\mathbf{B}}}_{\mathbf{I}_k}\boldsymbol{\Sigma}_{\mathbf{B}}^{-1}\mathbf{U}_{\mathbf{B}}^{\mathrm{T}} = \mathbf{U}_{\mathbf{B}}\underbrace{\boldsymbol{\Sigma}_{\mathbf{B}}\boldsymbol{\Sigma}_{\mathbf{B}}^{-1}}_{\mathbf{I}_k}\mathbf{U}_{\mathbf{B}}^{\mathrm{T}} = \mathbf{I}_k$.

To bound the second term above, we drop $\mathbf{V}_k^{\mathrm{T}}$, add and subtract $\mathbf{A}_{\rho-k}\mathbf{R}(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\mathrm{T}}$, and use the triangle inequality and spectral sub-multiplicativity,

$$\begin{aligned}&\|\mathbf{A}_{\rho-k}\mathbf{R}(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\dagger}\mathbf{V}_k^{\mathrm{T}}\|_{\mathrm{F}}\\ &\leq \|\mathbf{A}_{\rho-k}\mathbf{R}(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\mathrm{T}}\|_{\mathrm{F}} + \|\mathbf{A}_{\rho-k}\mathbf{R}((\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\dagger} - (\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\mathrm{T}})\|_{\mathrm{F}}\\ &\leq \|\mathbf{A}_{\rho-k}\mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{V}_k\|_{\mathrm{F}} + \|\mathbf{A}_{\rho-k}\mathbf{R}\|_{\mathrm{F}}\|(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\dagger} - (\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\mathrm{T}}\|_2.\end{aligned}$$

Now, we will bound each term individually. We bound the first term using $\mathcal{E}_3$. The second term can be bounded using $\mathcal{E}_1$ and $\mathcal{E}_2$ together with Lemma 8 (set $\mathbf{Q} = \mathbf{V}_k$ and $\Theta = \mathbf{R}$). Hence,

$$\begin{aligned}\|\widetilde{\mathbf{E}}\|_{\mathrm{F}} &\leq \|\mathbf{A}_{\rho-k}\mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{V}_k\|_{\mathrm{F}} + \|\mathbf{A}_{\rho-k}\mathbf{R}\|_{\mathrm{F}}\|(\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\dagger} - (\mathbf{V}_k^{\mathrm{T}}\mathbf{R})^{\mathrm{T}}\|_2\\ &\leq \varepsilon\|\mathbf{A}_{\rho-k}\|_{\mathrm{F}} + \sqrt{(1+\varepsilon)}\|\mathbf{A}_{\rho-k}\|_{\mathrm{F}}\cdot 1.5\varepsilon\\ &\leq \varepsilon\|\mathbf{A}_{\rho-k}\|_{\mathrm{F}} + 2\varepsilon\|\mathbf{A}_{\rho-k}\|_{\mathrm{F}}\\ &= 3\varepsilon\cdot\|\mathbf{A}_{\rho-k}\|_{\mathrm{F}}.\end{aligned}$$

The last inequality holds by our choice of $\varepsilon \in (0, 1/3)$. ■

Proof of Eqn. (2): $\mathbb{E}\,\|\mathbf{E}\|_{\mathrm{F}}^2 \leq (1+\varepsilon)\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2 \rightarrow \mathbb{E}\,\|\mathbf{E}\|_{\mathrm{F}}^2 - \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2 \leq \varepsilon\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2$. Now, apply Markov's inequality on the random variable $Y = \|\mathbf{E}\|_{\mathrm{F}}^2 - \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2 \geq 0$. ($Y \geq 0$ because $\mathbf{E} = \mathbf{A} - \mathbf{AZZ}^{\mathrm{T}}$ and $\mathrm{rank}(\mathbf{AZZ}^{\mathrm{T}}) = k$). This gives $\|\mathbf{E}\|_{\mathrm{F}}^2 - \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2 \leq 100\varepsilon\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2$ w.p. 0.99; so, $\|\mathbf{E}\|_{\mathrm{F}}^2 \leq \|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2 + 100\varepsilon\|\mathbf{A}-\mathbf{A}_k\|_{\mathrm{F}}^2$. ■

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments on a preliminary version of the manuscript.

REFERENCES

[1] J. A. Hartigan, *Clustering Algorithms*. New York, NY, USA: Wiley, 1975.

[2] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[3] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy, "The effectiveness of Lloyd-type methods for the k-means problem," in *Proc. 47th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2006, pp. 165–176.

[4] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.

[5] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror, "Result analysis of the NIPS 2003 feature selection challenge," in *Neural Information Processing Systems*. Red Hook, NY, USA: Curran & Associates Inc., 2005.

[6] R. Ostrovsky and Y. Rabani, "Polynomial time approximation schemes for geometric k-clustering," in *Proc. 41st Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, 2000, pp. 349–358.

[7] A. Kumar, Y. Sabharwal, and S. Sen, "A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions," in *Proc. 45th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, 2004, pp. 454–462.

[8] S. Har-Peled and S. Mazumdar, "On coresets for k-means and k-median clustering," in *Proc. 36th Annu. ACM Symp. Theory Comput. (STOC)*, 2004, pp. 291–300.

[9] S. Har-Peled and A. Kushal, "Smaller coresets for k-median and k-means clustering," in *Proc. 21st Annu. Symp. Comput. Geometry (SoCG)*, 2005, pp. 126–134.

[10] G. Frahling and C. Sohler, "A fast k-means implementation using coresets," in *Proc. 22nd Annu. Symp. Comput. Geometry (SoCG)*, 2006, pp. 135–143.

[11] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2007, pp. 1027–1035.

[12] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay, "Clustering in large graphs and matrices," in *Proc. 10th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 1999, pp. 291–299.
[13] D. Feldman, M. Schmidt, and C. Sohler, "Turning big data into tiny data: Constant-size coresets for *k*-means, PCA and projective clustering," in *Proc. 24th Annu. ACM-SIAM SODA*, 2013, pp. 1434–1453.
[14] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.
[15] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemp. Math.*, vol. 26, pp. 189–206, 1984.
[16] C. Boutsidis, M. W. Mahoney, and P. Drineas, "Unsupervised feature selection for the *k*-means clustering problem," in *Neural Information Processing Systems*. Red Hook, NY, USA: Curran & Associates Inc., 2009.
[17] M. Rudelson and R. Vershynin, "Sampling from large matrices: An approach through geometric functional analysis," *J. ACM*, vol. 54, no. 4, 2007, Art. ID 21.
[18] T. Sarlos, "Improved approximation algorithms for large matrices via random projections," in *Proc. 47th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2006, pp. 143–152.
[19] C. Boutsidis, P. Drineas, and M. Magdon-Ismail. (2011). "Near optimal column based matrix reconstruction." [Online]. Available: http://arxiv.org/abs/1103.0995
[20] M. D. Vose, "A linear algorithm for generating random numbers with a given distribution," *IEEE Trans. Softw. Eng.*, vol. 17, no. 9, pp. 972–975, Sep. 1991.
[21] M. Magdon-Ismail. (2010). "Row sampling for matrix algorithms via a non-commutative bernstein bound." [Online]. Available: http://arxiv.org/abs/1008.0587
[22] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proc. 30th Annu. ACM Symp. Theory Comput. (STOC)*, 1998, pp. 604–613.
[23] N. Ailon and B. Chazelle, "Approximate nearest neighbors and the fast Johnson–Lindenstrauss transform," in *Proc. 38th Annu. ACM Symp. Theory Comput. (STOC)*, 2006, pp. 557–563.
[24] D. Achlioptas, "Database-friendly random projections: Johnson–Lindenstrauss with binary coins," *J. Comput. Syst. Sci.*, vol. 66, no. 4, pp. 671–687, 2003.
[25] E. Liberty and S. W. Zucker, "The Mailman algorithm: A note on matrix–vector multiplication," *Inf. Process. Lett.*, vol. 109, no. 3, pp. 179–182, 2009.
[26] *MATLAB, 7.13.0.564 (R2011b)*, MathWorks, Natick, MA, USA, 2010.
[27] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Neural Information Processing Systems*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Red Hook, NY, USA: Curran & Associates Inc., 2006, pp. 507–514.
[28] *Feature Ranking Using Laplacian Score*. [Online]. Available: http://www.cad.zju.edu.cn/home/dengcai/Data/MCFS.html, accessed Jun. 4, 2013.
[29] K. Bache and M. Lichman. (2013). "UCI machine learning repository," School Inf. Comput. Sci., Univ. California, Irvine, CA, USA, Tech. Rep. [Online]. Available: http://archive.ics.uci.edu/ml
[30] S. A. Nene, S. K. Nayar, and H. Murase. (Feb. 1996). "Columbia University image library," Tech. Rep. CUCS-005-96. [Online]. Available: http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php
[31] AT&T Lab., Cambridge, U.K. *The ORL Database of Faces*. [Online]. Available: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html, accessed Jan 1, 2013.
[32] Carnegie Mellon Univ. *Pie Database*. [Online]. Available: http://www.ri.cmu.edu/research_project_detail.html?project_id=418&menu_id=261, accessed Jan 1, 2013.
[33] P. Drineas, R. Kannan, and M. Mahoney, "Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication," *SIAM J. Comput.*, vol. 36, no. 1, pp. 132–157, 2006.
[34] K. L. Clarkson, "Tighter bounds for random projections of manifolds," in *Proc. 24th Annu. Symp. Comput. Geometry (SoCG)*, 2008, pp. 39–48.
[35] S. Arora, E. Hazan, and S. Kale, "A fast random sampling algorithm for sparsifying matrices," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* (Lecture Notes in Computer Science), vol. 4110. Berlin, Germany: Springer-Verlag, 2006, pp. 272–279. [Online]. Available: http://dx.doi.org/10.1007/11830924_26

Christos Boutsidis is a Research Scientist at Yahoo Labs in New York, NY. Before that he was a Research Staff Member at the Business Analytics and Mathematical Sciences Department of the IBM T.J. Watson Research Center in Yorktown Heights, NY. Dr Boutsidis earned a Ph.D. in Computer Science from Rensselaer Polytechnic Institute in May of 2011 and a BS in Computer Engineering and Informatics from the University of Patras, Greece in July of 2006. Dr Boutsidis' research interests lie in the design and analysis of fast approximation algorithms for matrix computations and applications of those to machine learning and data analysis problems. Dr. Boutsidis has published over 25 articles in conferences and journals in numerical linear algebra, theoretical computer science, and statistical data analysis.

Anastasios Zouzias received his M.Sc. degree and Ph.D. degree in computer science from the University of Toronto, Canada, in 2009 and 2013, respectively. Currently, he has been with IBM Research Zurich, Switzerland, where he is a post-doctoral research scientist. His research interests include information retrieval, machine learning, randomized approximation algorithms, and randomized algorithms for numerical linear algebra.

Michael W. Mahoney is at the University of California at Berkeley in the Department of Statistics and at the International Computer Science Institute. He works on algorithmic and statistical aspects of modern large-scale data analysis. Much of his recent research has focused on large-scale machine learning, including randomized matrix algorithms and randomized numerical linear algebra, geometric network analysis tools for structure extraction in large informatics graphs, scalable implicit regularization methods, and applications in genetics, astronomy, medical imaging, social network analysis, and internet data analysis. He received him PhD from Yale University with a dissertation in computational statistical mechanics, and he has worked and taught at Yale University in the mathematics department, at Yahoo Research, and at Stanford University in the mathematics department. Among other things, he is on the national advisory committee of the Statistical and Applied Mathematical Sciences Institute (SAMSI), he was on the National Research Council's Committee on the Analysis of Massive Data, he runs the biennial MMDS Workshops on Algorithms for Modern Massive Data Sets, and he spent fall 2013 at UC Berkeley co-organizing the Simons Foundation's program on the Theoretical Foundations of Big Data Analysis.

Petros Drineas is an Associate Professor at the Computer Science Department of Rensselaer Polytechnic Institute. Prof. Drineas earned a PhD in Computer Science from Yale University in May of 2003, and a BS in Computer Engineering and Informatics from the University of Patras, Greece, in July of 1997. Prof. Drineas' research interests lie in the design and analysis of randomized algorithms for linear algebraic problems, as well as their applications to the analysis of modern, massive datasets. Prof. Drineas was a Visiting Professor at the US Sandia National Laboratories during the fall of 2005, a Visiting Fellow at the Institute for Pure and Applied Mathematics at the University of California, Los Angeles in the fall of 2007, and a Visiting Professor at the University of California Berkeley in the fall of 2013. Prof. Drineas has also served the US National Science Foundation (NSF) as a Program Director in the Information and Intelligent Systems (IIS) Division and the Computing and Communication Foundations (CCF) Division (2010–2011). Prof. Drineas has published over 90 articles in conferences and journals in Theoretical Computer Science, Numerical Linear Algebra, and statistical data analysis.