Optimization of the K-means Clustering Algorithm through Initialized Principal Direction Divisive Partitioning

Bruce James

Mentor: Dr. Jacob Kogan

July 23, 2017

University of Maryland, Baltimore County

Abstract

Data clustering is invaluable to the automated analysis of large document sets. Documents are converted into vectors in a finite dimensional space, and the resulting collection of salient features is then processed through an algorithm of one's choice, such as the classic k-means clustering algorithm. Due to the size of the feature space, different algorithms offer a trade-off between accuracy and computational efficiency. This study investigates the Principal Direction Divisive Partitioning (PDDP) algorithm, described as a top-down hierarchical technique, as a plug-in to the k-means algorithm. K-means reliance on initial random partitioning builds computational cost into the analysis. Using a PDDP initialized partition to seed k-means, computational efficiency will be compared to a k-means trial without PDDP.

Introduction

Clustering techniques have found numerous uses within the disparate fields of research that rely on data analysis. A cursory list of beneficiaries includes data mining, genomics, bioinformatics, and signal processing, with the field of text mining providing an historic impetus in its own right. Since the advent of the World Wide Web, cluster analysis has further developed in tandem with the rapidly increasing volume of document data existing online. Where human-based classification of text is assisted by an unsupervised approach, the organization and retrieval of documents from large data sets invariably becomes an automated task. With the proliferation of clustering algorithms to handle different text mining problems, the cost-benefit analysis of using different classification schemes has become a deciding factor in which method to use.

More than just providing motivation to develop existing methods, the field of document analysis also requires the development of the mathematical tools of optimization to increase the utility of such methods. Here, the linear algebra and optimization that supports clustering algorithms take a center position in the discussion. Naturally, data that is typically catalogued, as in text-based data, lends itself to the problem of finding the "optimal" clustering.

The clustering problem has been defined as the task of partitioning data with high in-group similarity, and low out-group similarity [11]. Implementing such a method with automation makes cluster analysis amenable to many applications in data mining and unsupervised machine learning. Yet, within the reach of clustering applications, many types of clustering algorithms have been developed, with none under as much scrutiny as the classic k-means algorithm.

The classic k-means algorithm is a partition-based, iterative process, using some distance-like function to establish a measure of similarity in the data [11]. Inadequacies in the results of k-means have become the reason for developing further algorithms. However, these developments have made k-means a benchmark for measuring the inspired algorithms.

In this paper, an alternative algorithm to k-means, the Principal Direction Divisive Partitioning (PDDP) clustering algorithm developed by Daniel Boley is described [2]. One limitation of k-means has been shown to arise from how the procedure is initialized [6]. Using a random initialization, k-means may prove to be computationally costly when compared to algorithms having a heuristic advantage. In contrast, PDDP has been shown to be an algorithm with exceptional utility in the clustering of document data.

The majority of this work appeals to conventional linear algebra and optimization techniques that underwrite the validity of the above-mentioned algorithms. First, the relevant notation and terminology are developed from the literature. The documents are then set in the context of high dimensional vector spaces, using a Euclidean norm to provide a distance-like measure, with the purpose of finding a least-squares approximation of the resulting data set to a "line of best fit." An objective function is defined and, using Lagrange multipliers, the least-squares approximation is reduced to an optimization problem. The relevant import, and necessary properties of the covariance matrix are then presented. This motivates the further discussions on the computation of the leading eigenvector, which makes up the principal direction used in Boley's original paper [1].

The ultimate goal of this investigation extends further to the coding of a working computer program, which will run the k-means algorithm with, and then without PDDP. The results of both trials will then be compared, where it is anticipated that the PDDP trial will run

more efficiently than the solo k-means trial. A demonstrable efficiency in the PDDP initialized program may suggest that initial directed partitioning of large text documents is superior to randomly seeded partitioning, in both computational runtime and accuracy of document clustering.

Literature Review

The character of many types of clustering methods depends on the applications involved. Document analysis differ from other applications like spatial recognition and robotic vision; sample spaces in which large document sets rely, do not necessarily possess the regularity of extended space [6]. From its human-generated origins, text typically contains irregularity, where the addition of new text evolves in unpredictable ways [4]. Historically, probabilistic methods have been used to anticipate this irregularity. Ultimately this ignores vast amounts of text, and greatly affects the accuracy of the analysis [2]. Contrasting with probabilistic means, the popular k-means algorithm is deterministic in its implementation, often with strategic use of random partitioning [2].

The k-means algorithm has long been an industry standard in partition-based clustering of texts. As such, much of the developed literature starts with the procedure's shortcomings when performing highly specialized tasks. Yet the algorithm has also been used as a benchmark in the literature for measuring the performance of many other boutique procedures. Here, a short description of the k-means algorithm will provide a wider coin of vantage to the survey of document clustering.

In the field of document analysis, mathematical and information theoretic tools are used to prepare semantically valued text for an objective treatment. First, document sets are grouped

together and relevant features, in the form of "content-bearing words" [4], are extracted and used to create a high-dimensional vector space. Individual documents would then exist as very sparse vectors (i.e. most word frequencies are zero, as any given document uses only a small fraction of the lexical set) in the vector space. Then, the k-means algorithm would provide a similarity-measure for clusters of documents having affinity within a group, compared to clusters outside of the group [11].

One computational consideration that arises when running k-means, originates from the random selection of centers used to initialize the algorithm. With no unique starting point, many resulting cluster arrangements may be found. The algorithm converges very fast, and even on large data sets this is not a problem. However, there is no guarantee that k-means will converge to the global minimum. Convergence of the algorithm to the global minimum is NP-hard. This exceeds realistic computational time constraints, and requires some additional heuristics to reduce computational effort [6].

As shown in Dhillon, et al. [4], one such heuristic, the spherical k-means technique, was expanded upon to work with the limitations in the k-means "hill climbing" strategy. After multiple iterations, both Euclidean and spherical k-means tends to become stuck in "qualitatively poor" local maxima, and not fully realize the optimal clustering scheme [4]. This is due to the gradient ascent scheme underwriting spherical k-means [3], and the problem of finding the "nearest cluster center" [5]. The latter issue stems from the discrete clustering problem (vs. a continuous problem that is readily accessible to linear algebra techniques [5]), where finding the global optimal clustering is NP-hard [5]. As the standard definition of optima is based on neighborhoods existing in a metric space, some work was needed to provide meaning to optimal partitioning. Drineas, et al. defined optimal clustering as a set of orthonormal vectors that has the

smallest possible "error" matrix, found by summing the squares of the matrix's entries after subtracting some set of clusters with negligible weight (Weight is the frequency of an occurrence of a node of a corresponding graph.) [5].

To then address optimization, Dhillon, et al. [4] suggested a "ping-pong" strategy that increases the computational efficiency of k-means by forming a sequence of separate clusters and moving certain documents from one cluster to another. The objective function could be studied to find a better optimization, thereby circumventing the optimization difficulty. As noted by the authors of that study, the approach was limited to static amounts of documents clustered. Further work would be required for large variances in sizes of the initial document set.

Optimization was extended further in Kogan, et al. [9] by developing the distance-like function from combining the Squared-Euclidean distance with an information theoretic quantity. It was suggested that the resulting similarity measures could be tailored to specific data sets. Whole classes of variations on the k-means theme could then be studied.

Limitations to the standard k-means algorithm have also been treated with various hybrid approaches [7], using an algorithm like PDDP to guide the k-means algorithm into trajectories with higher quality. The principal direction divisive partitioning (PDDP) algorithm, under consideration in this study, was developed by Daniel Boley. The method used a "divisive" method, where initially large document sets were divided into smaller partitions [2]. Boley used the term "principal direction" for the foremost direction that is computed at the beginning of all iterations dividing the document set. It was shown that this approach further increasing computational efficiency [2].

The mathematical treatment implemented in the PDDP algorithm can be summarized as follows. The projection of a set of vectors onto the nearest line starts the principal direction.

From there, the problem reverts to maximizing the leading eigenvalue of the covariance matrix, by using the power method [8]. Here, exploitation of the symmetric properties of the covariance matrix suggests the use of singular valued decomposition (SVD). According to Boley, increased accuracy of the computation can be achieved through the use of SVD [1]. Boley then implements the Lanczos algorithm for rapid computation of a partial SVD. Applying these linear algebra techniques significantly reduces computational complexity and run-time [1], thereby rendering PDDP more practical for use in high-dimensional cluster analysis.

Research Objective

The ultimate objective of this study is to investigate a hybrid application of the PDDP algorithm to initialize the k-means clustering algorithm, as a way to reduce the computational effort exerted by the stand-alone k-means algorithm. The reduction in computational cost would compensate for the irregularity of text data, and enable more accuracy for sorting and querying from large document sets. A comprehensive review of the optimization performed on the PDDP algorithm, as well as cluster validation measures, will inform further investigations on the quality of the resulting clusters, and suggest further work on the strengths and limitations of hybrid approaches.

Analysis of the PDDP Algorithm

High-dimensional vector spaces, equipped with some distance-like function, are the standard environment for large document sets. This study begins with an overview of the linear algebra and optimization techniques, which underwrite these spaces. With deference to the notation of the prior works, vectors representing documents with *m* features will be represented

as boldfaced, lowercase letters, namely, $\mathbf{b} = [b_1, b_2, ..., b_m]^T$, and a collection of n documents will be denoted by the matrix $B = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n]$.

We start by finding the projection, $\mathbf{p}_a$ on a line L in $\mathbf{R}^n$, parameterized by $\mathbf{y} + t\,\mathbf{x}$, on which is projected the document vector $\mathbf{a}$. Note that the vector $(\mathbf{y} - t_0\,\mathbf{x}) - \mathbf{a}$ is orthogonal to L, and also the vector $\mathbf{x}$ at the point of projection, for some t_0 , namely

$$(\mathbf{y} + t_0\,\mathbf{x} - \mathbf{a})^T\,\mathbf{x} = 0 \;\Rightarrow\; \mathbf{y}^T\mathbf{x} + t_0\,\|\mathbf{x}\|^2 - \mathbf{a}^T\mathbf{x} = 0$$

$$\Rightarrow\; t_0 = (\mathbf{a}^T\mathbf{x} - \mathbf{y}^T\mathbf{x})/\,\|\mathbf{x}\|^2.$$

So, $\mathbf{p}_a = \mathbf{y} + t_0\,\mathbf{x}$, where $t_0 = (\mathbf{a}^T\mathbf{x} - \mathbf{y}^T\mathbf{x})/\,\|\mathbf{x}\|^2$, which is the point that $\mathbf{a}$ projects on L. With this, then $(\mathbf{a} - \mathbf{p}_a)^T\mathbf{x} = 0$. If we take $\mathbf{x}$ to lie on the unit sphere, and $\mathbf{y}$ orthogonal to $\mathbf{x}$, such that $\mathbf{x}^T\mathbf{x} = 1 \Rightarrow \|\mathbf{x}\|^2 = 1$, and $\mathbf{y}^T\mathbf{x} = 0$, then

$$\mathbf{p}_a = \mathbf{y} + [(\mathbf{a}^T\mathbf{x} - \mathbf{y}^T\mathbf{x})/\,\|\mathbf{x}\|^2]\,\mathbf{x}$$

$$= \mathbf{y} + (\mathbf{a}^T\mathbf{x})\,\mathbf{x}.$$

In this form, we hope to find the least squares approximation of a set of document vectors. To do this, we will minimize the sum of the distances of each projection $\mathbf{p}_i$ on L, with $\mathbf{a}_i$. In terms of $\mathbf{x}^T\mathbf{x} = 1$,and $\mathbf{y}^T\mathbf{x} = 0$**,** the sum, then is

$$\Sigma_{1\le i\le n}\;|\mathbf{a}_i - (\mathbf{y} + (a_i{}^T x)\,\mathbf{x})|^2.$$

Regrouping and simplifying each term $|\mathbf{a} - \mathbf{x}\mathbf{a}^T\mathbf{x} - \mathbf{y}|^2$ of the sum, gives

$$|(\mathbf{a} - \mathbf{x}\mathbf{a}^T\mathbf{x}) - \mathbf{y}|^2 = (\mathbf{a} - \mathbf{x}\mathbf{a}^T\mathbf{x})^2 - 2\mathbf{y}(\mathbf{a} - \mathbf{x}\mathbf{a}^T\mathbf{x}) + \mathbf{y}^2$$

$$= \mathbf{a}\cdot\mathbf{a} - 2\mathbf{a}(\mathbf{x}\mathbf{a}^T\mathbf{x}) + (\mathbf{x}\mathbf{a}^T\mathbf{x})^2 - 2\mathbf{a}\cdot\mathbf{y} + 2\mathbf{y}\cdot\mathbf{x}\mathbf{a}^T\mathbf{x} + \mathbf{y}\cdot\mathbf{y}$$

$$= \mathbf{a}\cdot\mathbf{a} - 2(\mathbf{a}\cdot\mathbf{x})(\mathbf{a}\cdot\mathbf{x}) + (\mathbf{x}\cdot\mathbf{x})(\mathbf{a}^T\mathbf{x})^2 - 2\mathbf{a}\cdot\mathbf{y} + 2(\mathbf{y}\cdot\mathbf{x})\mathbf{a}^T\mathbf{x} + \mathbf{y}\cdot\mathbf{y}$$

$$= \mathbf{a}\cdot\mathbf{a} - 2(\mathbf{a}^T\mathbf{x})^2 + (1)(\mathbf{a}^T\mathbf{x})^2 - 2\mathbf{a}\cdot\mathbf{y} + 2(\mathbf{0})\mathbf{a}^T\mathbf{x} + \mathbf{y}\cdot\mathbf{y}$$

$$= \mathbf{a}\cdot\mathbf{a} - (\mathbf{a}^T\mathbf{x})^2 - 2\mathbf{a}^T\mathbf{y} + \mathbf{y}\cdot\mathbf{y}$$

$$= (\mathbf{a} - \mathbf{y})^2 - (\mathbf{a}^T\mathbf{x})^2.$$

So,

$$\Sigma_{1\le i\le n}\ |\mathbf{a}_i - (\mathbf{y} + (a_i^T x)\ \mathbf{x})|^2 = \Sigma_{1\le i\le n}\ |\mathbf{a}_i - \mathbf{y}|^2 - |\mathbf{a}_i{}^{\mathrm{T}}\mathbf{x}|^2.$$

This sum is a "fit" line of the data set.

We now want a "best fit" line. Assuming a known value for **x**, we use the method of Lagrange multipliers to find **y**, such that

$$\min_{\mathbf{y}} \left\{\Sigma_{1\le i\le n}\ |\mathbf{a}_i - \mathbf{y}|^2, \text{ subject to } \mathbf{y}^{\mathrm{T}}\mathbf{x} = 0\right\}. \qquad (1)$$

Let $f(\mathbf{y})$ be the vector-valued objective function $f(y_1,\ldots,y_n) = \Sigma_{1\le i\le n}\ |\mathbf{a}_i - \mathbf{y}|^2$, and $g(\mathbf{y}) = \mathbf{y}^{\mathrm{T}}\mathbf{x} = 0$ be the constraint function. Taking the gradient of f yields

$$\nabla f(\mathbf{y}) = \Sigma_{1\le j\le n}\ f_j(y_1,\ldots,y_n)\ \mathbf{e}_j,$$

where f_j is the partial derivative of f with respect to the j-th component, and $\mathbf{e}_j$ is the unit basis vector of the j-th component, namely $\mathbf{e}_j = [0,\ldots, \mathbf{e}_j = 1,\ldots,0]^{\mathrm{T}}$.

With $|\mathbf{a}_i - \mathbf{y}|^2 = \mathbf{a}_i{}^{\mathrm{T}}\mathbf{a}_i - 2\mathbf{a}_i{}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y}$ for each i, then

$$\nabla f(\mathbf{y}) = \Sigma_{1\le j\le n}\ \partial/\partial y_j\ \left[\Sigma_{1\le i\le n}\ |\mathbf{a}_i - \mathbf{y}|^2\right]\mathbf{e}_j$$

$$= \Sigma_{1\le j\le n}\ \partial/\partial y_j\ \left[\Sigma_{1\le i\le n}\ \mathbf{a}_i{}^{\mathrm{T}}\mathbf{a}_i - 2\mathbf{a}_i{}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y}\right]\mathbf{e}_j$$

Since we are in $\mathbf{R}^n$, and the 2-norm is an assignment from $\mathbf{R}^n$ to $\mathbf{R}$ ($|\mathbf{a}_i - \mathbf{y}|^2$ is the Euclidean norm assigning the n-vector **y** to a real number), then from a theorem in real analysis, f is continuous. So, from another property of analysis, we can interchange the summand with partial derivative, yielding

$$\nabla f(\mathbf{y}) = \Sigma_{1\le j\le n}\ \Sigma_{1\le i\le n}\ \partial/\partial y_j\ \left[\mathbf{a}_i{}^{\mathrm{T}}\mathbf{a}_i - 2\mathbf{a}_i{}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y}\right]\mathbf{e}_j$$

$$= \Sigma_{1\le j\le n}\ \Sigma_{1\le i\le n}\ \partial/\partial y_j\ \left[\mathbf{a}_i{}^{\mathrm{T}}\mathbf{a}_i - 2(a_{ji}y_j) + y_j^2\right]\mathbf{e}_j$$

$$= \Sigma_{1\le j\le n}\ \left[\Sigma_{1\le i\le n}\ (-2a_{ji} + 2y_j)\ \mathbf{e}_j\right]$$

$$= \Sigma_{1\le j\le n}\left[\Sigma_{1\le i\le n}(-a_{ji})\,\mathbf{e}_j + \Sigma_{1\le i\le n} y_j\,\mathbf{e}_j\right]$$

$$= 2\left[\Sigma_{1\le i\le n}(-\mathbf{a}_i) + \Sigma_{1\le i\le n}\mathbf{y}\right]$$

$$= 2n\mathbf{y} - 2\,\Sigma_{1\le i\le n}\,\mathbf{a}_i.$$

Now, taking the gradient of g, gives

$$\nabla g(\mathbf{y}) = \Sigma_{1\le j\le n}\,\partial/\partial y_j\,[y_j x_j]\,\mathbf{e}_j$$

$$\nabla g(\mathbf{y}) = \partial/\partial y_j\,[y_1x_1 + \ldots + y_nx_n]\mathbf{e}_j$$

$$= \Sigma_{1\le j\le n}\,(x_j)\mathbf{e}_j$$

$$= \mathbf{x}.$$

Employing the Lagrange multiplier, and solving the system for lambda yields

$$\nabla f(\mathbf{y}) = \lambda\nabla g(\mathbf{y}) \quad \Rightarrow \quad 2n\mathbf{y} - 2\,\Sigma_{1\le i\le n}\,\mathbf{a}_i = \lambda\,\mathbf{x}$$

Now, left-multiplying by $\mathbf{x}^T$, and using the conditions $\mathbf{y}^T\mathbf{x} = 0 \Rightarrow \mathbf{x}^T\mathbf{y} = 0$, and $\mathbf{x}^T\mathbf{x} = 1$,

$$\mathbf{x}^T\,(2n\mathbf{y} - 2\,\Sigma_{1\le i\le n}\,\mathbf{a}_i) = \mathbf{x}^T\,\lambda\,\mathbf{x}$$

$$\Rightarrow \quad \mathbf{x}^T\,(2n)\mathbf{y} - \mathbf{x}^T\,(2)\,\Sigma_{1\le i\le n}\,\mathbf{a}_i = \lambda\,(\mathbf{x}^T\,\mathbf{x})$$

$$\Rightarrow \quad (2n)\mathbf{x}^T\mathbf{y} - 2\mathbf{x}^T\,\Sigma_{1\le i\le n}\,\mathbf{a}_i = \lambda\,(\mathbf{x}^T\,\mathbf{x})$$

$$\Rightarrow \quad \lambda = -2\mathbf{x}^T\,\Sigma_{1\le i\le n}\,\mathbf{a}_i.$$

Substituting for lambda gives

$$2n\mathbf{y} - 2\,\Sigma_{1\le i\le n}\,\mathbf{a}_i = -2\mathbf{x}^T\left(\Sigma_{1\le i\le n}\,\mathbf{a}_i\right)\mathbf{x}$$

$$\Rightarrow \quad \mathbf{y} = 1/n\left[\Sigma_{1\le i\le n}\,\mathbf{a}_i - \mathbf{x}^T\left(\Sigma_{1\le i\le n}\,\mathbf{a}_i\right)\mathbf{x}\right],$$

with $\mathbf{x}$ and its transpose as known entities. So, to minimize (2),

$$\mathbf{y} = 1/n \left[\Sigma_{1\le i\le n}\, \mathbf{a}_i \;-\; \mathbf{x}^{\mathrm{T}} \left(\Sigma_{1\le i\le n}\, \mathbf{a}_i\right)\mathbf{x}\right].$$

Now, we turn to the collection of column vectors representing the document set. First, we note the following property:

Property 1:

For $B = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n]$, with $A = \mathbf{b}_1\mathbf{b}_1^{\mathrm{T}} + \mathbf{b}_2\mathbf{b}_2^{\mathrm{T}} + \ldots + \mathbf{b}_n\mathbf{b}_n^{\mathrm{T}}$, then $BB^{\mathrm{T}} = A$.

Proof:

Since matrix B is an $m \times n$ array and B^{T} is an $n \times m$ array, then BB^{T} is an $m \times m$ array. Similarly for A, each $\mathbf{b_k}\mathbf{b_k}^{\mathbf{T}}$ term, where

$$\mathbf{b_k}\mathbf{b_k}^{\mathrm{T}} = [b_{\mathrm{k}1}\ b_{\mathrm{k}2}\ \ldots\ b_{\mathrm{km}}\ ]^{\mathrm{T}}\ [b_{\mathrm{k}1}\ b_{\mathrm{k}2}\ \ldots\ b_{\mathrm{km}}\ ],$$

is an $m \times m$ array, whose sum for $1 \le k \le n$ is also an $m \times m$ array. Let the ij^{th} entry of BB^{T} be denoted by c_{ij}. Then, performing matrix multiplication yields,

$$c_{ij} = b_{i1}b_{1j} + b_{i2}b_{2j} + \ldots + b_{im}b_{mj}.$$

Let the ij^{th} entry of $\mathbf{b_k}\mathbf{b_k}^{\mathrm{T}} = d^{(k)}_{ij}$. Carrying out matrix multiplication gives $d^{(k)}_{ij} = b_{ik}b_{kj}$. Performing addition for $1 \le k \le n$ yields,

$$\Sigma_{1\le k\le n}\, d^{(k)}_{ij} = b_{i1}b_{1j} + b_{i2}b_{2j} + \ldots + b_{im}b_{mj}.$$

This sum is the ij^{th} entry of matrix A. Since $c_{ij} = \Sigma_{1\le k\le n}\, d^{(k)}_{ij}$ for the ij^{th} entries of BB^{T} and A respectively, therefore,

$$BB^{\mathrm{T}} = A.$$

Now, we want to exploit the symmetric and positive semi-definite properties of the matrix A.

Property 2:

A is positive semi-definite.

Proof:

From the associative properties of matrix multiplication,

$$\mathbf{x}^T\mathbf{B}\ \mathbf{B}^T\mathbf{x} = (\mathbf{x}^T\mathbf{B})(\mathbf{B}^T\mathbf{x}).$$

Since $\mathbf{x}^T\mathbf{B} = (\mathbf{B}^T\mathbf{x})^T$, and $\mathbf{x}^T\mathbf{B}$ is the transpose of a column vector $\mathbf{B}^T\mathbf{x}$, then $(\mathbf{x}^T\mathbf{B})(\mathbf{B}^T\mathbf{x})$ is a dot product of $\mathbf{B}^T\mathbf{x}$ with itself. From the positivity property of inner product spaces (i.e. $\langle v,v\rangle \geq 0$, for all $v \in \mathbf{R}^n$), then

$$\mathbf{x}^T\mathbf{B}\ \mathbf{B}^T\mathbf{x} \geq 0.$$

So, *A* is positive semi-definite.

With the fact that *A* is positive semi-definite, then we also know that all of the eigenvalues of *A* are real. We gather this from the real spectral theorem, and since **A** can be decomposed into the product $\mathbf{UDU}^T$, where **U** is an orthogonal matrix and **D** is diagonal, with real entries on the main diagonal equal to the eigenvalues of **A**. Hence, the eigenvalues are real [10].

We state the next property.

Property 3:

Every eigenvalue of A is non-negative.

Proof:

Since **A** is symmetric, then for any arbitrary eigenvalue λ of **A** satisfying $\mathbf{Ax} = \lambda\mathbf{x}$, $\lambda \in \mathbf{R}$, left multiplying $\mathbf{x}^T$ to both sides of $\mathbf{Ax} = \lambda\mathbf{x}$, yields

$$\mathbf{x}^T\mathbf{Ax} = \mathbf{x}^T\lambda\mathbf{x} = \lambda\mathbf{x}^T\mathbf{x}.$$

Since **A** is positive semi-definite,

$$\mathbf{x}^T\mathbf{Ax} \geq 0 \ \Rightarrow\ \lambda\mathbf{x}^T\mathbf{x} \geq 0,$$

and by positivity of inner product spaces,

$$\mathbf{x}^T\mathbf{x} \geq 0 \;\; \Rightarrow \;\; \lambda \geq 0.$$

Therefore, the eigenvalues of A are non-negative.

After discussing some of the useful properties of the covariance matrix, which comprises matrix multiplication of the document column vectors with their transpose, it is now desired to find the least squares approximation of the "fit" line of the data set. We start by substituting $\mathbf{c} - \mathbf{x}(\mathbf{c}^T\mathbf{x})$ in for $\mathbf{y}$ to find

$$\min_{\mathbf{x},\mathbf{y}} \left\{ \sum_{1 \leq i \leq n} |\mathbf{a}_i - \mathbf{y}|^2 - |\mathbf{x}^T\mathbf{a}_i|^2, \text{ subject to } \mathbf{y}^T\mathbf{x} = 0, \mathbf{x}^T\mathbf{x} = 1 \right\}, \tag{2}$$

and simplifying the expression $|\mathbf{a}_i - \mathbf{y}|^2 - |\mathbf{x}^T\mathbf{a}_i|$, as such:

$$\begin{aligned}
|\mathbf{a}_i - \mathbf{y}|^2 - |\mathbf{x}^T\mathbf{a}_i| &= [(\mathbf{a}_i - \mathbf{c}) + \mathbf{x}(\mathbf{c}^T\mathbf{x})]^T[(\mathbf{a}_i - \mathbf{c}) + \mathbf{x}(\mathbf{c}^T\mathbf{x})] - |\mathbf{x}^T\mathbf{a}_i|^2 \\
&= |\mathbf{a}_i - \mathbf{c}|^2 + 2(\mathbf{a}_i - \mathbf{c})^T\, \mathbf{x}(\mathbf{c}^T\mathbf{x}) \\
&\qquad + ((\mathbf{c}^T\mathbf{x})^2\,)\mathbf{x}^T\mathbf{x} - (\mathbf{x}^T\mathbf{a}_i)^T(\mathbf{x}^T\mathbf{a}_i) \\
&= |\mathbf{a}_i - \mathbf{c}|^2 - [-2(\mathbf{a}_i - \mathbf{c})^T\, \mathbf{x}(\mathbf{c}^T\mathbf{x}) \\
&\qquad - ((\mathbf{c}^T\mathbf{x})^2\,)\mathbf{x}^T\mathbf{x} + (\mathbf{x}^T\mathbf{a}_i)^T(\mathbf{x}^T\mathbf{a}_i)] \\
&= |\mathbf{a}_i - \mathbf{c}|^2 - [-2\mathbf{a}_i{}^T\mathbf{x}(\mathbf{c}^T\mathbf{c}) + 2\mathbf{c}^T\mathbf{x}(\mathbf{c}^T\mathbf{c}) \\
&\qquad - (\mathbf{c}^T\mathbf{x})^2 + (\mathbf{x}^T\mathbf{a}_i)^T(\mathbf{x}^T\mathbf{a}_i)] \\
&= |\mathbf{a}_i - \mathbf{c}|^2 - [-2\mathbf{a}_i{}^T\mathbf{x}(\mathbf{c}^T\mathbf{c}) + (\mathbf{c}^T\mathbf{x})^2 + (\mathbf{x}^T\mathbf{a}_i)^2] \\
&= |\mathbf{a}_i - \mathbf{c}|^2 - |\mathbf{c}^T\mathbf{x} - \mathbf{a}_i{}^T\mathbf{x}|^2.
\end{aligned}$$

Since $|\mathbf{a}_i - \mathbf{c}|^2 \geq 0$, and the only expression that depends on $\mathbf{x}$ is

$$|\mathbf{c}^T\mathbf{x} - \mathbf{a}_i{}^T\mathbf{x}|^2 \geq 0, \; \forall \; \mathbf{x} \in \mathbf{R}^n,$$

then the minimum of (1) is the largest value of $(\mathbf{c}^T\mathbf{x} - \mathbf{a}_i{}^T\mathbf{x})^2$, namely

$$\sup_{\mathbf{x}} \left\{ \sum_{1 \leq i \leq n} |\mathbf{c}^T\mathbf{x} - \mathbf{a}_i{}^T\mathbf{x}|^2, \text{ constrained to } \mathbf{x}^T\mathbf{x} = 1 \right\}. \tag{3}$$

In order to set-up the optimization problem, let $f(\mathbf{x})$ be the vector-valued objective function

$$f(x_1,\ldots,x_n) = \Sigma_{1\le i\le n}\ \left[|\mathbf{c}^{\mathrm{T}}\mathbf{x} - \mathbf{a}_i{}^{\mathrm{T}}\mathbf{x}|^2\right],$$

and

$$g(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{x}$$

be the constraint function. Taking the gradient of f yields

$$\nabla f(\mathbf{x}) = \Sigma_{1\le j\le n}\ f_j(x_1,\ldots,x_n)\ \mathbf{e}_j,$$

where f_j is the partial derivative of f with respect to the j-th component, and $\mathbf{e}_j$ is the unit basis vector of the j-th component namely $\mathbf{e}_j=[0,\ldots,e_j=1,\ldots,0]^{\mathrm{T}}$.

Then

$$\nabla f(\mathbf{x}) = \Sigma_{1\le j\le n}\ \partial/\partial x_j\ \left[\ \Sigma_{1\le i\le n}\ \ \left[|\mathbf{c}^{\mathrm{T}}\mathbf{x} - \mathbf{a}_i{}^{\mathrm{T}}\mathbf{x}|^2\right]\right]\mathbf{e}_j$$

$$= \Sigma_{1\le j\le n}\Sigma_{1\le i\le n}\left[\ \partial/\partial x_j\ \left[|\mathbf{c}^{\mathrm{T}}\mathbf{x} - \mathbf{a}_i{}^{\mathrm{T}}\mathbf{x}|^2\right]\right]\mathbf{e}_j$$

$$= \Sigma_{1\le j\le n}\Sigma_{1\le i\le n}\left[\ \partial/\partial x_j\ \left[|(\mathbf{c} - \mathbf{a}_i)^{\mathrm{T}}\mathbf{x}|^2\right]\right]\mathbf{e}_j$$

$$= \Sigma_{1\le j\le n}\Sigma_{1\le i\le n}\left[\ \partial/\partial x_j\ \left[((\mathbf{c} - \mathbf{a}_i)^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}((\mathbf{c} - \mathbf{a}_i)^{\mathrm{T}}\mathbf{x})\right]\right]\mathbf{e}_j$$

$$= \Sigma_{1\le j\le n}\Sigma_{1\le i\le n}\left[\ \partial/\partial x_j\ \left[\mathbf{x}^{\mathrm{T}}(\mathbf{c} - \mathbf{a}_i)(\mathbf{c} - \mathbf{a}_i)^{\mathrm{T}}\mathbf{x})\right]\right]\mathbf{e}_j$$

$$= \Sigma_{1\le j\le n}\Sigma_{1\le i\le n}\left[\ \partial/\partial x_j\ \left[\mathbf{x}^{\mathrm{T}}\left((\mathbf{c} - \mathbf{a}_i)(\mathbf{c} - \mathbf{a}_i)^{\mathrm{T}}\right)\mathbf{x}\right]\mathbf{e}_j.\right.$$

Using the results from assignment 2, as follows:

Letting $\mathbf{c} - \mathbf{a}_i = \mathbf{b}_i$, then

$$\Sigma_{1\le j\le n}\Sigma_{1\le i\le n}\left[\ \partial/\partial x_j\ \left[\mathbf{x}^{\mathrm{T}}\left((\mathbf{c} - \mathbf{a}_i)(\mathbf{c} - \mathbf{a}_i)^{\mathrm{T}}\right)\mathbf{x}\right]\mathbf{e}_j = \Sigma_{1\le j\le n}\Sigma_{1\le i\le n}\left[\partial/\partial x_j\ \mathbf{x}^{\mathrm{T}}\left(\mathbf{b}_i\mathbf{b}_i{}^{\mathrm{T}}\right)\mathbf{x}\right]\mathbf{e}_j.\right.$$

From the results in problem 1(b), assignment 2, $BB^{\mathrm{T}} = \sum_{1 \le i \le n} \mathbf{b}_i \mathbf{b}_i^{\mathrm{T}}$.

So,

$$\sum_{1 \le j \le n} \sum_{1 \le i \le n} \left[\partial / \partial x_j \ \mathbf{x}^{\mathrm{T}} \left(\mathbf{b}_i \mathbf{b}_i^{\mathrm{T}} \right) \mathbf{x} \right] \mathbf{e}_j = \sum_{1 \le j \le n} \partial / \partial x_j \left[\mathbf{x}^{\mathrm{T}} BB^{\mathrm{T}} \mathbf{x} \right] \mathbf{e}_j .$$

The result from problem 1(c), assignment 2 also states that $\mathbf{x}^{\mathrm{T}} BB^{\mathrm{T}} \mathbf{x}$ is symmetric, and positive semi-definite. Therefore, all of the eigenvalues of BB^{T} are real.

Now, taking the gradient of *g,* gives

$$\nabla g(\mathbf{x}) = \sum_{1 \le j \le n} \partial / \partial x_j \left[\mathbf{x}^{\mathrm{T}} \mathbf{x} \right] \mathbf{e}_j ,$$

where $\mathbf{x}^{\mathrm{T}} \mathbf{x} = 1$. Employing the Lagrange multiplier, and solving the system for lambda yields

$$\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x})$$

$$\Rightarrow \ \sum_{1 \le j \le n} \partial / \partial x_j \left[\mathbf{x}^{\mathrm{T}} BB^{\mathrm{T}} \mathbf{x} \right] \mathbf{e}_j = \lambda \sum_{1 \le j \le n} \partial / \partial x_j \left[\mathbf{x}^{\mathrm{T}} \mathbf{x} \right] \mathbf{e}_j .$$

$$\Rightarrow \int \sum_{1 \le j \le n} \partial / \partial x_j \left[\mathbf{x}^{\mathrm{T}} BB^{\mathrm{T}} \mathbf{x} \right] \mathbf{e}_j \, dx_j = \int \lambda \sum_{1 \le j \le n} \partial / \partial x_j \left[\mathbf{x}^{\mathrm{T}} \mathbf{x} \right] \mathbf{e}_j \, dx_j$$

$$\Rightarrow \sum_{1 \le j \le n} \int \partial / \partial x_j \left[\mathbf{x}^{\mathrm{T}} BB^{\mathrm{T}} \mathbf{x} \right] \mathbf{e}_j \, dx_j = \lambda \sum_{1 \le j \le n} \int \partial / \partial x_j \left[\mathbf{x}^{\mathrm{T}} \mathbf{x} \right] \mathbf{e}_j \, dx_j$$

$$\Rightarrow \mathbf{x}^{\mathrm{T}} BB^{\mathrm{T}} \mathbf{x} = \lambda \mathbf{x}^{\mathrm{T}} \mathbf{x} = \lambda .$$

Squaring both sides gives

$$(\mathbf{x}^{\mathrm{T}} BB^{\mathrm{T}} \mathbf{x})^2 = (\lambda \mathbf{x})^2 \ \Rightarrow \ (\mathbf{x}^{\mathrm{T}} BB^{\mathrm{T}} \mathbf{x})^{\mathrm{T}} (\mathbf{x}^{\mathrm{T}} BB^{\mathrm{T}} \mathbf{x}) = \lambda^2 \, \mathbf{x}^{\mathrm{T}} \mathbf{x}$$

$$\Rightarrow \ |\mathbf{x}^{\mathrm{T}} BB^{\mathrm{T}} \mathbf{x}|^2 = \lambda^2$$

$$\Rightarrow \ \lambda = |\mathbf{x}^{\mathrm{T}} BB^{\mathrm{T}} \mathbf{x}|$$

Note that $\mathbf{x}\mathbf{x}^{\mathrm{T}}$ is an n x n array, such that

Left multiplying by **x,** and noting that $\mathbf{x}^{\mathrm{T}} \mathbf{x} = 1 \Rightarrow \mathbf{x}\mathbf{x}^{\mathrm{T}} = 1$, yields

$$(\mathbf{x}\mathbf{x}^{\mathrm{T}}) BB^{\mathrm{T}} \mathbf{x} = \lambda (\mathbf{x}\mathbf{x}^{\mathrm{T}}) \mathbf{x} \ \Rightarrow \ BB^{\mathrm{T}} \mathbf{x} = \lambda \mathbf{x},$$

where **x** is the eigenvector to some real eigenvalue, lambda. To find lambda, left multiply the last equation by $\mathbf{x}^T$ to get

$$\mathbf{x}^T BB^T\mathbf{x} = \mathbf{x}^T\lambda\mathbf{x} = \lambda(\mathbf{x}^T\mathbf{x}) = \lambda.$$

So, to minimize (3), **x** is an eigenvector, with the corresponding real eigenvalue $\lambda = \mathbf{x}^T BB^T\mathbf{x}$, of the matrix BB^T.

Extensions and Future Work

Mathematical Analysis

Further mathematical treatment of the PDDP algorithm will look into finding the leading eigenvector obtained through the optimization process. The SVD and power methods used to increase efficiency and accuracy of computation, including the Lanczos algorithm, will also be studied. Investigation of the use of the Frobenius norm of the covariance matrix would also provide a segue into the latter steps in the completion of PDDP [2].

Experimental Exploration

The latter work of this study hopes explores two programs running the PDDP and k-means algorithms over a set of document data. The algorithms will be written in Python. Documents would then be procured from web-based sources, such as MEDLINE, for sample research abstracts and larger documents. An off-the-shelf porter stemmer will be required to reduce the morphological complexity of the texts.

The clustered documents will come from various disciplines, with optimal variance. Two sets of trials will be made, with one trial set consisting of a collection of research abstracts, and another trial set from the bodies of the research papers. Each document set will then be run

through a porter stemmer, and the resulting morphologically reduced data set fed into the PDDP initialized k-means program, and again, through the solo k-means program.

Computational Analysis

Under further investigation, it would be of interest to measure the efficiency and accuracy of the experimental clusters. Computational run-time will be measured in Python. In keeping with prior research, clustering validation measures are employed to determine the quality of the final clusters; both internal and external validation techniques would be useful here. The accuracy of each trial will be determined through a confusion matrix [9]. Anticipating a uniform effect from the back-end of both k-means applications (with and without PDDP), the clusters are qualitatively assessed by comparing the "true" class size with the experimental results. Since entropy-based assessments of the resulting clusters may not adequately measure uniform effects, a Coefficient of Variation statistic can here be applied to the trial clusters [12].

References

[1] Berry, M. W., Dumais, S. T., and O'Brien, G. W. Using linear algebra for intelligent information retrieval. *SIAM Review*, (4), 573, 1995.

[2] Boley, D. Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, *2*(4), 325–34, 1998.

[3] Dhillon, I., Dharmendra, M. Concept decompositions for large sparse text data using clustering. *Machine Learning, 42*(1-2), 143-175, 2001.

[4] Dhillon, I. S., Guan, Y., and Kogan, J. Iterative clustering of high dimensional text data augmented by local search. In *IEEE International Conference on Data Mining, 2002. Proceedings.* 131–138, 2002.

[5] Drineas, P., Frieze, A., Kannan, R., Vempala, S., & Vinay, V. Clustering in large graphs and matrices. *Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 291-299). Society for Industrial and Applied Mathematics, 1999.

[6] Hamerly, G., Elkan, C., Kalpakis, K., Goharian, N., and Grossman, D. Alternatives to the k-means algorithm that find better clusterings. In K. Kalpakis, N. Goharian, and D. Grossman (Eds.). Presented at the Proceedings of the Eleventh International Conference on Information and Knowledge Management. CIKM, 2002.

[7] Kogan, J. Hybrid clustering of large text data. In *21st International Conference on Advanced Information Networking and Applications Workshops, 1*, 2007.

[8] Kogan, J. *Introduction to clustering large and high-dimensional data*. Cambridge New York: Cambridge University Press, 2007.

[9] Kogan, J., Teboulle, M., and Nicholas, C. Data driven similarity measures for k means like clustering algorithms. *Information Retrieval, (8)* 331-349, 2005.

[10] Larson, R. *Elementary linear algebra, 7th Ed*. Boston Massachusetts: Brooks/Cole, 2013.

[11] Reddy, C., Aggarwal, C. *Data clustering: Algorihms and applications.* Boca Raton Florida: CRC Press, 2014

[12] Wu, J. *Advances in K-means clustering. [electronic resource] : a data mining thinking*. Berlin ; New York : Springer, pages 17-35. 2012.