K-means Clustering via Principal Component Analysis

Chris Ding CHQDING@LBL.GOV
Xiaofeng He XHE@LBL.GOV
Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Abstract

Principal component analysis (PCA) is a widely used statistical technique for unsupervised dimension reduction. K-means clustering is a commonly used data clustering for unsupervised learning tasks. Here we prove that principal components are the continuous solutions to the discrete cluster membership indicators for K-means clustering. Equivalently, we show that the subspace spanned by the cluster centroids are given by spectral expansion of the data covariance matrix truncated at $K-1$ terms. These results indicate that unsupervised dimension reduction is closely related to unsupervised learning. On dimension reduction, the result provides new insights to the observed effectiveness of PCA-based data reductions, beyond the conventional noise-reduction explanation. Mapping data points into a higher dimensional space via kernels, we show that solution for Kernel K-means is given by Kernel PCA. On learning, our results suggest effective techniques for K-means clustering. DNA gene expression and Internet newsgroups are analyzed to illustrate the results. Experiments indicate that newly derived lower bounds for K-means objective are within 0.5-1.5% of the optimal values.

1. Introduction

Data analysis methods are essential for analyzing the ever-growing massive quantity of high dimensional data. On one end, cluster analysis(Duda et al., 2000; Hastie et al., 2001; Jain & Dubes, 1988) attempts to pass through data quickly to gain first order knowledge by partitioning data points into disjoint groups such that data points belonging to same cluster are similar while data points belonging to different clusters are dissimilar. One of the most popular and efficient clustering methods is the K-means method (Hartigan & Wang, 1979; Lloyd, 1957; MacQueen, 1967) which uses prototypes (centroids) to represent clusters by optimizing the squared error function. (A detail account of K-means and related ISODATA methods are given in (Jain & Dubes, 1988), see also (Wallace, 1989).)

On the other end, high dimensional data are often transformed into lower dimensional data via the principal component analysis (PCA)(Jolliffe, 2002) (or singular value decomposition) where coherent patterns can be detected more clearly. Such unsupervised dimension reduction is used in very broad areas such as meteorology, image processing, genomic analysis, and information retrieval. It is also common that PCA is used to project data to a lower dimensional subspace and K-means is then applied in the subspace (Zha et al., 2002). In other cases, data are embedded in a low-dimensional space such as the eigenspace of the graph Laplacian, and K-means is then applied (Ng et al., 2001).

The main basis of PCA-based dimension reduction is that PCA picks up the dimensions with the largest variances. Mathematically, this is equivalent to finding the best low rank approximation (in L_2 norm) of the data via the singular value decomposition (SVD) (Eckart & Young, 1936). However, this noise reduction property alone is inadequate to explain the effectiveness of PCA.

In this paper , we explore the connection between these two widely used methods. We prove that principal components are actually the continuous solution of the cluster membership indicators in the K-means clustering method, i.e., the PCA dimension reduction automatically performs data clustering according to the K-means objective function. This provides an important justification of PCA-based data reduction.

Our results also provide effective ways to solve the K-

Appearing in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

means clustering problem. K-means method uses K prototypes, the centroids of clusters, to characterize the data. They are determined by minimizing the sum of squared errors,

$$J_{\rm K} = \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \mathbf{m}_k)^2$$

where $(\mathbf{x}_1, \cdots, \mathbf{x}_n) = X$ is the data matrix and $\mathbf{m}_k = \sum_{i \in C_k} \mathbf{x}_i / n_k$ is the centroid of cluster C_k and n_k is the number of points in C_k. Standard iterative solution to K-means suffers from a well-known problem: as iteration proceeds, the solutions are trapped in the local minima due to the greedy nature of the update algorithm (Bradley & Fayyad, 1998; Grim et al., 1998; Moore, 1998).

Some notations on PCA. X represents the original data matrix; $Y = (\mathbf{y}_1, \cdots, \mathbf{y}_n)$, $\mathbf{y}_i = \mathbf{x}_i - \bar{\mathbf{x}}$, represents the centered data matrix, where $\bar{\mathbf{x}} = \sum_i \mathbf{x}_i / n$. The covarance matrix (ignoring the factor $1/n$) is $\sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = YY^T$. Principal directions $\mathbf{u}_k$ and principal components $\mathbf{v}_k$ are eigenvectors satisfying:

$$YY^T \mathbf{u}_k = \lambda_k \mathbf{u}_k,\ Y^T Y \mathbf{v}_k = \lambda_k \mathbf{v}_k,\ \mathbf{v}_k = Y^T \mathbf{u}_k / \lambda_k^{1/2}. \tag{1}$$

These are the defining equations for the SVD of Y: $Y = \sum_k \lambda_k^{1/2} \mathbf{u}_k \mathbf{v}_k^T$ (Golub & Van Loan, 1996). Elements of $\mathbf{v}_k$ are the projected values of data points on the principal direction $\mathbf{u}_k$.

2. 2-way clustering

Consider the $K = 2$ case first. Let

$$d(\mathrm{C}_k, \mathrm{C}_\ell) \equiv \sum_{i \in C_k} \sum_{j \in C_\ell} (\mathbf{x}_i - \mathbf{x}_j)^2$$

be the sum of squared distances between two clusters $\mathrm{C}_k, \mathrm{C}_\ell$. After some algebra we obtain

$$J_{\rm K} = \sum_{k=1}^{K} \sum_{i,j \in C_k} \frac{(\mathbf{x}_i - \mathbf{x}_j)^2}{2n_k} = n\overline{\mathbf{y}^2} - \frac{1}{2} J_{\rm D}, \tag{2}$$

and

$$J_{\rm D} = \frac{n_1 n_2}{n} \left[2\frac{d(\mathrm{C}_1, \mathrm{C}_2)}{n_1 n_2} - \frac{d(\mathrm{C}_1, \mathrm{C}_1)}{n_1^2} - \frac{d(\mathrm{C}_2, \mathrm{C}_2)}{n_2^2} \right] \tag{3}$$

where $\overline{\mathbf{y}^2} = \sum_i \mathbf{y}_i^T \mathbf{y}_i / n$ is a constant. Thus $\min(J_{\rm K})$ is equivalent to $\max(J_{\rm D})$. Furthermore, we can show

$$\frac{d(\mathrm{C}_1, \mathrm{C}_2)}{n_1 n_2} = \frac{d(\mathrm{C}_1, \mathrm{C}_1)}{n_1^2} + \frac{d(\mathrm{C}_2, \mathrm{C}_2)}{n_2^2} + (\mathbf{m}_1 - \mathbf{m}_2)^2. \tag{4}$$

Substituting Eq.(4) into Eq.(3), we see $J_{\rm D}$ is always positive. We summarize these results in

Theorem 2.1. For $K = 2$, minimization of K-means cluster objective function $J_{\rm K}$ is equivalent to maximization of the distance objective $J_{\rm D}$, which is always positive.

Remarks. (1) In $J_{\rm D}$, the first term represents average between-cluster distances which are maximized; this forces the resulting clusters as separated as possible. (2) The second and third terms represent the average within-cluster distances which will be minimized; this forces the resulting clusters as compact or tight as possible. This is also evident from Eq.(2). (3) The factor $n_1 n_2$ encourages cluster balance. Since $J_{\rm D} > 0$, $\max(J_{\rm D})$ implies maximization of $n_1 n_2$, which leads to $n_1 = n_2 = n/2$.

These remarks give some insights to the K-means clustering. However, the primary importance of Theorem 2.1 is that $J_{\rm D}$ leads to a solution via the principal component.

Theorem 2.2. For K-means clustering where $K = 2$, the continuous solution of the cluster indicator vector is the principal component $\mathbf{v}_1$, i.e., clusters C_1, C_2 are given by

$$C_1 = \{i \mid \mathbf{v}_1(i) \le 0\}, \quad C_2 = \{i \mid \mathbf{v}_1(i) > 0\}. \tag{5}$$

The optimal value of the K-means objective satisfies the bounds

$$n\overline{\mathbf{y}^2} - \lambda_1 < J_{\rm K=2} < n\overline{\mathbf{y}^2} \tag{6}$$

Proof. Consider the squared distance matrix $D = (d_{ij})$, where $d_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||^2$. Let the cluster indicator vector be

$$q(i) = \begin{cases} \sqrt{n_2/nn_1} & \text{if} \quad i \in \mathrm{C}_1 \\ -\sqrt{n_1/nn_2} & \text{if} \quad i \in \mathrm{C}_2 \end{cases} \tag{7}$$

This indicator vector satisfies the sum-to-zero and normalization conditions: $\sum_i q(i) = 0$, $\sum_i q^2(i) = 1$. One can easily see that $\mathbf{q}^T D \mathbf{q} = -J_{\rm D}$. If we relax the restriction that $\mathbf{q}$ must take one of the two discrete values, and let $\mathbf{q}$ take any values in $[-1, 1]$, the solution of minimization of $J(\mathbf{q}) = \mathbf{q}^T D \mathbf{q} / \mathbf{q}^T \mathbf{q}$ is given by the eigenvector corresponding to the lowest (largest negative) eigenvalue of the equation $D\mathbf{z} = \lambda \mathbf{z}$.

A better *relaxation* of the discrete-valued indicator $\mathbf{q}$ into continuous solution is to use the centered distance matrix D, i.e, to subtract column and row means of D. Let $\hat{D} = (\hat{d}_{ij})$, where

$$\hat{d}_{ij} = d_{ij} - d_{i.}/n - d_{.j}/n + d_{..}/n^2 \tag{8}$$

where $d_{i\cdot} = \sum_j d_{ij}, d_{\cdot j} = \sum_i d_{ij}, d_{\cdot\cdot} = \sum_{ij} d_{ij}$. Now we have $\mathbf{q}^T \hat{D}\mathbf{q} = \mathbf{q}^T D\mathbf{q} = -J_{\rm D}$, since the 2nd, 3rd and 4th terms in Eq.(8) contribute zero in $\mathbf{q}^T\hat{D}\mathbf{q}$. Therefore the desired cluster indicator vector is the eigenvector corresponding to the lowest (largest negative) eigenvalue of

$$\hat{D}\mathbf{z} = \lambda\mathbf{z}.$$

By construction, this centered distance matrix $\hat{D}$ has a nice property that each row (and column) is sum-to-zero, $\sum_i \hat{d}_{ij} = 0,\ \forall j$. Thus $\mathbf{e} = (1, \cdots, 1)^T$ is an eigenvector of $\hat{D}$ with eigenvalue $\lambda = 0$. Since all other eigenvectors of $\hat{D}$ are orthogonal to $\mathbf{e}$, i.e, $\mathbf{z}^T\mathbf{e} = 0$, they have the sum-to-zero property, $\sum_i \mathbf{z}(i) = 0$, a definitive property of the initial indicator vector $\mathbf{q}$. In contrast, eigenvectors of $D\mathbf{z} = \lambda\mathbf{z}$ do not have this property.

With some algebra, $d_{i\cdot} = n\mathbf{x}_i^2 + n\overline{\mathbf{x}^2} - 2n\mathbf{x}_i^T\bar{\mathbf{x}},\ d_{\cdot\cdot} = 2n^2\overline{\mathbf{y}^2}$. Substituting into Eq.(8), we obtain

$$\hat{d}_{ij} = -2(\mathbf{x}_i - \bar{\mathbf{x}})^T(\mathbf{x}_j - \bar{\mathbf{x}}) \quad \text{or} \quad \hat{D} = -2Y^TY.$$

Therefore, the continuous solution for cluster indicator vector is the eigenvector corresponding to the largest (positive) eigenvalue of the Gram matrix Y^TY, which by definition, is precisely the principal component $\mathbf{v}_1$. Clearly, $J_{\rm D} < 2\lambda_1$, where λ_1 is the principal eigenvalue of the covariance matrix. Through Eq.(2), we obtain the bound on $J_{\rm K}$. □

Figure 1 illustrates how the principal component can determine the cluster memberships in K-means clustering. Once C_1, C_2 are determined via the principal component according to Eq.(5), we can compute the current cluster means $\mathbf{m}_k$ and iterate the K-means until convergence. This will bring the cluster solution to the local optimum. We will call this PCA-guided K-means clustering.

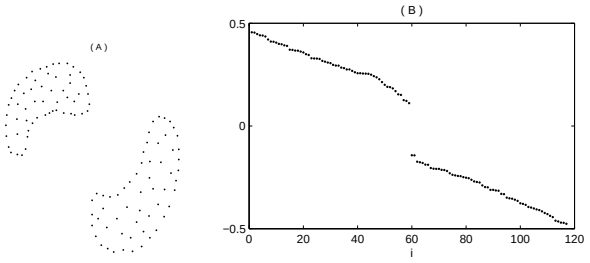

Figure 1. (A) Two clusters in 2D space. (B) Principal component $\mathbf{v}_1(i)$, showing the value of each element i.

3. K-way Clustering

Above we focus on the $K = 2$ case using a single indicator vector. Here we generalize to $K > 2$, using $K-1$ indicator vectors.

Regularized relaxation

This general approach is first proposed in (Zha et al., 2002). Here we present a much expanded and consistent relaxation scheme and a connectivity analysis. First, with the help of Eq.(2), $J_{\rm K}$ can be written as

$$J_{\rm K} = \sum_i \mathbf{x}_i^2 - \sum_k \frac{1}{n_k} \sum_{i,j\in C_k} \mathbf{x}_i^T\mathbf{x}_j, \tag{9}$$

The first term is a constant. The second term is the sum of the K diagonal block elements of X^TX matrix representing within-cluster (inner-product) similarities.

The solution of the clustering is represented by K non-negative indicator vectors: $H_K = (\mathbf{h}_1, \cdots, \mathbf{h}_K)$, where

$$\mathbf{h}_k = (0, \cdots, 0, \overbrace{1, \cdots, 1}^{n_k}, 0, \cdots, 0)^T / n_k^{1/2} \tag{10}$$

(Without loss of generality, we index the data such that data points within each cluster are adjacent.) With this, Eq.(9) becomes

$$J_{\rm K} = {\rm Tr}(X^TX) - {\rm Tr}(H_k^TX^TXH_k) \tag{11}$$

where ${\rm Tr}(H_K^TX^TXH_K) = \mathbf{h}_1^TX^TX\mathbf{h}_1 + \cdots + \mathbf{h}_k^TX^TX\mathbf{h}_k$. There are redundancies in H_K. For example, $\sum_{k=1}^K n_k^{1/2}\mathbf{h}_k = \mathbf{e}$. Thus one of the $\mathbf{h}_k$'s is linear combination of others. We remove this redundancy by (a) performing a linear transformation T into $\mathbf{q}_k$'s:

$$Q_K = (\mathbf{q}_1, \cdots, \mathbf{q}_K) = H_KT, \text{ or } \mathbf{q}_\ell = \sum_k \mathbf{h}_k t_{k\ell}, \tag{12}$$

where $T = (t_{ij})$ is a $K \times K$ orthonormal matrix: $T^TT = I$, and (b) requiring that the last column of T is

$$\mathbf{t}_n = (\sqrt{n_1/n}, \cdots, \sqrt{n_k/n})^T. \tag{13}$$

Therefore we always have

$$\mathbf{q}_K = \sqrt{\frac{n_1}{n}}\mathbf{h}_1 + \cdots + \sqrt{\frac{n_k}{n}}\mathbf{h}_K = \sqrt{\frac{1}{n}}\mathbf{e}.$$

This linear transformation is always possible (see later). For example when $K = 2$, we have

$$T = \begin{pmatrix} \sqrt{n_2/n} & -\sqrt{n_1/n} \\ \sqrt{n_1/n} & \sqrt{n_2/n} \end{pmatrix}, \tag{14}$$

and $\mathbf{q}_1 = \sqrt{n_2/n}\,\mathbf{h}_1 - \sqrt{n_1/n}\,\mathbf{h}_2$, which is precisely the indicator vector of Eq.(7). This approach for K-way clustering is the generalization of $K = 2$ clustering in §2.

The mutual orthogonality of $\mathbf{h}_k$, $\mathbf{h}_k^T\mathbf{h}_\ell = \delta_{k\ell}$ ($\delta_{k\ell} = 1$ if $k = \ell$; 0 otherwise), implies the mutual orthogonality of $\mathbf{q}_k$,

$$\mathbf{q}_k^T\mathbf{q}_\ell = \sum_p \mathbf{h}_p^T t_{pk} \sum_s \mathbf{h}_s t_{s\ell} = \sum_p (T^T T)_{k\ell} = \delta_{k\ell}.$$

Let $Q_{K-1} = (\mathbf{q}_1, \cdots, \mathbf{q}_{K-1})$, the above orthogonality relation can be represented as

$$Q_{K-1}^T Q_{K-1} = I_{K-1}, \tag{15}$$

$$\mathbf{q}_k^T\mathbf{e} = 0, \text{for } k = 1, \cdots, K-1. \tag{16}$$

Now, the K-means objective can be written as

$$J_{\rm K} = \mathrm{Tr}(X^T X) - \mathbf{e}^T X^T X\mathbf{e}/n - \mathrm{Tr}(Q_{k-1}^T X^T X Q_{k-1}) \tag{17}$$

Note that $J_{\rm K}$ does not distinguish the original data $\{\mathbf{x}_i\}$ and the centered data $\{\mathbf{y}_i\}$. Repeating the above derivation on $\{\mathbf{y}_i\}$, we have

$$J_{\rm K} = \mathrm{Tr}(Y^T Y) - \mathrm{Tr}(Q_{k-1}^T Y^T Y Q_{k-1}), \tag{18}$$

noting that $Y\mathbf{e} = 0$ because rows of Y are centered. The first term is constant. Optimization of $J_{\rm K}$ becomes

$$\max_{Q_{K-1}} \mathrm{Tr}(Q_{K-1}^T Y^T Y Q_{K-1}) \tag{19}$$

subject to the constraints Eqs.(15,16), with additional constraint that $\mathbf{q}_k$ are the linear transformations of the $\mathbf{h}_k$ as in Eq.(12). If we relax (ignore) the last constraint, i.e., let $\mathbf{h}_k$ to take continuous values, while still keeping constraints Eqs.(15,16), the maximization problem can be solved in closed form, with the following results:

Theorem 3.1. When optimizing the K-means objective function, the continuous solutions for the transformed discrete cluster membership indicator vectors Q_{K-1} are the $K-1$ principal components: $Q_{K-1} = (\mathbf{v}_1, \cdots, \mathbf{v}_{K-1})$. $J_{\rm K}$ satisfies the upper and lower bounds

$$n\overline{\mathbf{y}^2} - \sum_{k=1}^{K-1} \lambda_k < J_{\rm K} < n\overline{\mathbf{y}^2} \tag{20}$$

where $n\overline{\mathbf{y}^2}$ is the total variance and λ_k are the principal eigenvalues of the covariance matrix YY^T.

Note that the constraints of Eq.(16) are automatically satisfied, because $\mathbf{e}$ is an eigenvector of $Y^T Y$ with $\lambda = 0$ and the orthogonality between eigenvectors associated with different eigenvalues. This result is true for any K. For $K = 2$, it reduces to that of §2.

The proof is a direct application of a well-known theorem of Ky Fan (Fan, 1949) (Theorem 3.2 below) to the optimization problem Eq.(19).

Theorem 3.2. (Fan) Let A be a symmetric matrix with eigenvalues $\zeta_1 \geq \cdots \geq \zeta_n$ and corresponding eigenvectors $(\mathbf{v}_1, \cdots, \mathbf{v}_n)$. The maximization of $\mathrm{Tr}(Q^T A Q)$ subject to constraints $Q^T Q = I_K$ has the solution $Q = (\mathbf{v}_1, \cdots, \mathbf{v}_K)R$, where R is an arbitary $K \times K$ orthonormal matrix, and $\max \mathrm{Tr}(Q^T A Q) = \zeta_1 + \cdots + \zeta_K$.

Eq.(11) is first noted in (Gordon & Henderson, 1977) in slghtly different form as a referee comment and was promptly dismissed. It is independently rediscovered in (Zha et al., 2002) where spectral relaxation technique is applied [to Eq.(11) instead of Eq.(18)], leading to K principal eigenvectors of $X^T X$ as the continuous solution. The present approach has three advantages:
(a) Direct relaxation on $\mathbf{h}_k$ in Eq.(11) is not as desirable as relaxation on $\mathbf{q}_k$ of Eq.(18). This is because $\mathbf{q}_k$ satisfy sum-to-zero property of the usual PCA components while $\mathbf{h}_k$ do not. Entries of discrete indicator vectors $\mathbf{q}_k$ have both positive and negative values, thus closer to the continuous solution. On the other hand, entries of discrete indicator vectors $\mathbf{h}_k$ have only one sign, while all eigenvectors (except $\mathbf{v}_1$) of $X^T X$ have both positive and negative entries. In other words, the continuous solutions of $\mathbf{h}_k$ will differ significantly from its discrete form, while the continuous solutions of $\mathbf{q}_k$ will be much closer to its discrete form.
(b) The present approach is consistent with both $K > 2$ case and $K = 2$ case presented in §2 using a single indicator vector. The relaxation of Eq.(11) for $K = 2$ would requires two eigenvectors; that is not consistent with the single indicator vector approach in §2.
(c) Relaxation in Eq.(11) uses the original data, $X^T X$, while the present approach uses centered matrix $Y^T Y$. Using $Y^T Y$ makes the orthogonality conditions Eqs.(15, 16) consistent since $\mathbf{e}$ is an eigenvector of $Y^T Y$. Also, $Y^T Y$ is closely related to covariance matrix YY^T, a central theme in statistics.

Cluster Subspace Identifcation

Suppose we have found K clusters with $\mathbf{m}_k$ centroids. The projection matrix $P = \sum_{k=1}^K \mathbf{m}_k\mathbf{m}_k^T$ project any vector $\mathbf{x}$ into the subspace spanned by the K centroids: $P^T\mathbf{x} = \sum_{k=1}^K (\mathbf{m}_k^T\mathbf{x})\mathbf{m}_k$ We call this subspace as cluster subspace. From Theorem 3.1, we have

Theorem 3.3. Cluster subspace is spanned by the first $K-1$ principal directions, i.e., $P = \sum_{k=1}^{K-1} \lambda_k \mathbf{u}_k\mathbf{u}_k^T$.
Proof. A cluster centroid $\mathbf{m}_k$ can be represented via the cluster indicator vector, $\mathbf{m}_k = \sum_{i\in C_k} \mathbf{y}_i = \sum_i h_k(i)\mathbf{y}_i = Yh_k$. Thus P becomes

$$\sum_{k=1}^K Y h_k h_k^T Y^T = Y \sum_{k=1}^K \mathbf{h}_k\mathbf{h}_k^T Y^T = Y \sum_{k=1}^K \mathbf{q}_k\mathbf{q}_k^T Y^T$$

Now, upon using Theorem 3.1, $\mathbf{q}_1, \mathbf{q}_{K-1}$ are given by $\mathbf{v}_1, \cdots, \mathbf{v}_{K-1}$, i.e., and $\mathbf{q}_K$ is given by $\mathbf{e}_1/n^{1/2}$. Thus $\sum_{k=1}^{K} \mathbf{q}_k \mathbf{q}_k^T \rightarrow \mathbf{e}\mathbf{e}^T/n + \sum_{k=1}^{K-1} \mathbf{v}_k \mathbf{v}_k^T$. Note $Y\mathbf{e} = 0$ because Y contained centered data. Using Eq.(1), $X\mathbf{v}_k = \lambda_k^{1/2} \mathbf{u}_k$. This completes the proof.

Theorem 3.3 implies that PCA dimension reduction automatically finds the cluster subspace. This fact explains why PCA dimension reduction is particularly beneficial for K-means clustering, because clustering in the cluster subspace is typically more effective than clustering in the original space, as explained in the following.

Proposition 3.4. In cluster subspace, between-cluster distances remain nearly as in original space, while within-cluster distances are reduced.

Proof. Writing the distance between $\mathbf{y}_i, \mathbf{y}_j$ as

$$||\mathbf{y}_i - \mathbf{y}_j||_d^2 = ||\mathbf{y}_i^\perp - \mathbf{y}_j^\perp||_r^2 + ||\mathbf{y}_i^\parallel - \mathbf{y}_j^\parallel||_s^2$$

where $\mathbf{y}_i^\parallel$ is the component in the r-dimensional cluster subspace and $\mathbf{y}_i^\perp$ is the component in the s-dimensional irrelavent subspace ($d = r + s$). We wish to show that

$$\frac{||\mathbf{y}_i^\parallel - \mathbf{y}_j^\parallel||_r}{||\mathbf{y}_i - \mathbf{y}_j||_d} \approx \begin{cases} 1 & \text{if } i \in C_k, j \in C_\ell \neq C_k \\ r/d & \text{if } i, j \in C_k, \end{cases} \tag{21}$$

If y_i, y_j are in different clusters, $\mathbf{y}_i - \mathbf{y}_j$ runs from one cluster to another, or, it runs from one centroid to another. Thus it is nearly inside the cluster subspace. This proves the first equality in Eq.21. If y_i, y_j are in the same cluster, we assume data has a Gaussian distribution. With probability of r/d, $\mathbf{y}_i - \mathbf{y}_j$ points to a direction in the cluster subspace, which are retained in after PCA projection. With probability of s/d, $\mathbf{y}_i - \mathbf{y}_j$ points to a direction outside the cluster subspace, which collaps to zero, $||\mathbf{y}_i^\perp - \mathbf{y}_j^\perp||^2 \approx 0$. This proves the second equality in Eq.21.

Eq.(21) shows that in cluster subspace, between-cluster distances remain constant; while within-cluster distances shrink: clusters become relatively more compact. The lower the cluster subspace dimension r is, the more compact clusters become, and the more effective the K-means clustering in the subspace.

When projecting to a subspace, the subspace representations could differ by an orthogonal transformation T. Because K-means clustering is invariant w.r.t. T, we do not need the explicit form of T.

In summary, the automatic identification of the cluster subspace via PCA dimension reduction guarrantees that K-means clustering in the PCA subspace is particularly effective.

Kernel K-means clustering and Kernel PCA

From Eq.(9), K-means clustering can be viwed as using the stand dot-product (Gram matrix). Thus it can be easily extended to any other kernels (Zhang & Rudnicky, 2002). This is done using a nonlinear transformation (a mapping) to the higher dimensional space

$$\mathbf{x}_i \rightarrow \phi(\mathbf{x}_i)$$

The clustering objective function under this mapping, with the help of Eq.(9), can be written as

$$\min J_{\mathrm{K}}(\phi) = \sum_i ||\phi(\mathbf{x}_i)||^2 - \sum_k \frac{1}{n_k} \sum_{i,j \in C_k} \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j), \tag{22}$$

The first term is a constant for a given mapping function $\phi(\cdot)$ and can be ignored. The objective function becomes

$$\max J_{\mathrm{K}}^W = \sum_k \frac{1}{n_k} \sum_{i,j \in C_k} w_{ij} \tag{23}$$

where $W = (w_{ij})$ is the kernel matrix: $w_{ij} = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$.

The advantage of Kernel K-means is that it can describe data distributions more complicated that Gaussion distributions. The disadvantage of Kernel K-means is that it no longer has cluster centroids because there are only pairwise kernel or similarities. Thus the fast order(n) local refinement no longer apply.

PCA has been applied to kernel matrix in (Schölkopf et al., 1998). Some advantages has been shown due to the nonlinear transformation. In light of the equivalence between K-means clustering and PCA shown above, the relationship between Kernel K-means clustering and Kernel PCA can be discerned easily. In fact, repeating previously analysis, we can show that solution to Kernel K-means is given by Kernel PCA components:

Theorem 3.5. The continuous solutions for the discrete cluster membership indicator vectors in Kernel K-means clustering are the $\kappa - 1$ Kernel PCA components: J_{K}^W satisfies the upper bounds

$$J_{\mathrm{K}}^W < \sum_{k=1}^{K-1} \zeta_k \tag{24}$$

where ζ_k are the principal eigenvalues of the Kernel PCA matrix W.

Recovering K clusters

Once the $K-1$ principal components $\mathbf{q}_k$ are computed, how to recover the non-negative cluster indicators $\mathbf{h}_k$, therefore the clusters themselves?

Clearly, since $\sum_i \mathbf{q}_k(i) = 0$, each principal component has many negative elements, so they differ substantially from the non-negative cluster indicators $\mathbf{h}_k$. Thus the key is to compute the orthonormal transformation T in Eq.(12).

A $K \times K$ orthonormal transformation is equivalent to a rotation in K-dimensional space; there are K^2 elements with $K(K+1)/2$ constraints (Goldstein, 1980); the remaining $K(K-1)/2$ degrees of freedom are most conveniently represented by Euler angles. For $K = 2$ a single rotation ϕ specifies the transformation; for $K = 3$ three Euler angles (ϕ, θ, ζ) determine the rotation, etc. In K-means problem, we require that the last column of T have the special form of Eq.(13); therefore, the true degree of freedom is $F_K = K(K-1)/2-1$. For $K = 2$, $F_K = 0$ and the solution is fixed; it is given in Eq.(14). For $K = 3$, $F_K = 2$ and we need to search through a 2-D space to find the best solution, i.e., to find the T matrix that will transform $\mathbf{q}_k$ to non-negative indicator vectors $\mathbf{h}_k$.

Using Euler angles to specify the orthogonal rotation in high dimensional space $K > 3$ with the special constraint is often complicated. This problem can be solved via the following representation. Given arbitrary $K(K-1)/2$ positive numbers α_{ij} that sum-to-one, $\sum_{1\le i\le j\le n} \alpha_{ij} = 1$ and $\alpha_{ij} = \alpha_{ji}$. The degree of freedom is $K(K-1)/2-1$, same as the degree of freedom in our problem. We form the following $K \times K$ matrix:

$$\mathit{\Gamma} = \Omega^{-1/2}\ \bar{\mathit{\Gamma}}\ \Omega^{-1/2},\ \Omega = \mathrm{diag}(\sqrt{n_1}, \cdots, \sqrt{n_K}).$$

where

$$\bar{\mathit{\Gamma}}_{ij} = -\alpha_{ij},\ i \neq j;\ \bar{\mathit{\Gamma}}_{ii} = \sum_{j|j\neq i} \alpha_{ij}, \qquad (25)$$

It can be shown that $2\sum_{ij} x_i \bar{\mathit{\Gamma}}_{ij} x_j = \sum_{ij} \alpha_{ij}(x_i - x_j)^2$ for any $\mathbf{x} = (x_1, \cdots, x_K)^T$. Thus the symmetric matrix $\mathit{\Gamma}$ is semi-positive definite. $\mathit{\Gamma}$ has K eigenvectors of real values with non-negative eigenvalues: $(\mathbf{z}_1, \cdots, \mathbf{z}_K) = Z$, where $\mathit{\Gamma}\mathbf{z}_k = \lambda_k \mathbf{z}_k$. Clearly, $\mathbf{t}_n$ of Eq.(13) is an eigenvector of $\mathit{\Gamma}$ with eigenvalue $\lambda = 0$. Other $K-1$ eigenvectors are mutually orthonormal, $Z^T Z = I$. Under general conditions, Z is non-singular and $Z^{-1} = Z^T$. Thus Z is the desired orthonormal transformation T in Eq.(12). Summerizing these results, we have

Theorem 3.3. The linear transformation T of Eq.(12) is formed by the K eigenvectors of $\mathit{\Gamma}$ specified by Eq.(25).

This result indicates the K-means clustering is reduced to an optimization problem with $K(K-1)/2-1$ parameters.

Connectivity Analysis

The above analysis gives the structure of the transformation T. But before knowning the clustering results, we cannot compute T and thus not computing $\mathbf{h}_k$ neither. Thus we need a method that bypasses T.

In Theorem 3.3, cluster subspace spanned via the cluster centroids is given by the first $K-1$ principal directions, i.e., the low-dimensional spectral representation of the covariance matrix YY^T. We examine the low-dimensional spectral representation of the kernel (Gram) matrix $Y^T Y = \sum_{k=1}^{K-1} \lambda_k \mathbf{v}_k \mathbf{v}_k^T$. For subspace projection, the coefficient λ_k is unimportant. Adding a constant matrix $\mathbf{e}\mathbf{e}^T/n^{1/2}$, we have

$$C = \mathbf{e}\mathbf{e}^T/n^{1/2} + \sum_{k=1}^{K-1} \mathbf{v}_k \mathbf{v}_k^T.$$

Follow the proof of Theorem 3.3, C can be written as

$$C = \sum_{k=1}^{K} \mathbf{q}_k \mathbf{q}_k^T = \sum_{k=1}^{K} \mathbf{h}_k \mathbf{h}_k^T.$$

$\sum_k \mathbf{h}_k \mathbf{h}_k^T$ has a clear diagonal block structure, which leads naturally to a connectivity interpretation: if $c_{ij} > 0$ then $\mathbf{x}_i, \mathbf{x}_j$ are in the same cluster, we say they are connected. We further associate a probability for the connectivity between i, j as $p_{ij} = c_{ij}/c_{ii}^{1/2} c_{jj}^{1/2} = \delta_{ij}$ depending whether $\mathbf{x}_i, \mathbf{x}_j$ are connected or not. The diagonal block structure is the characteristic structure for clusters.

If the data has clear cluster structure, we expect C has similar diagonal block structure, plus some noise, due to the fact that principal components are approximations of the discrete valued indicators. For example, C could contain negative elements. Thus we set $c_{ij} = 0$ if $c_{ij} < 0$. Also, elements in C with small positive values indicate weak, possibly spurious, connectivities, which should be suppressed. We set

$$c_{ij} = 0 \quad \text{if} \quad p_{ij} < \beta, \qquad (26)$$

where $0 < \beta < 1$ and we chose $\beta = 0.5$.

Once C is computed, the block structure can be detected using the spectral ordering(Ding & He, 2004). By computing the cluster crossing, the cluster overlap

along the specified ordering, a 1-D curve exhibits the cluster structure. Clusters can be identified using a linearized cluster assignment (Ding & He, 2004).

4. Experiments

Gene expressions

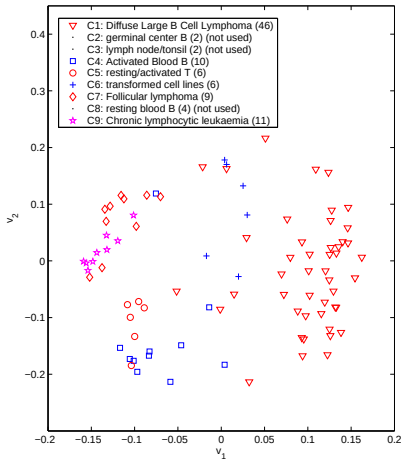

Figure 2. Gene expression profiles of human lymphoma(Alizadeh et al., 2000) in first two principal components.

4029 gene expressions of 96 tissue samples on human lymphoma is obtained by Alizadeh et al.(Alizadeh et al., 2000). Using biological and clinic expertise, Alizadeh et al classify the tissue samples into 9 classes as shown in Figure 2. Because of the large number of classes and also highly uneven number of samples in each classes (46, 2, 2, 10, 6, 6, 9, 4, 11), it is a relatively difficult clustering problem. To reduce dimension, 200 out of 4029 genes are selected based on F-statistic for this study. We focus on 6 largest classes with at least 6 tissue samples per class to adequately represent each class; classes C2, C3, and C8 are ignored because the number of samples in these classes are too small (8 tissue samples total). Using PCA, we plot the samples in the first two principal components as in Fig.2.

Following Theorem 3.1, the cluster structure are embedded in the first $K-1=5$ principal components. In this 5-dimensional eigenspace we perform K-means clustering. The clustering results are given in the following confusion matrix

$$B = \begin{bmatrix} 36 & \cdot & \cdot & \cdot & \cdot & \cdot \\ 2 & 10 & \cdot & \cdot & \cdot & 1 \\ 1 & \cdot & 9 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 11 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 6 & \cdot \\ 7 & \cdot & \cdot & \cdot & \cdot & 5 \end{bmatrix}$$

where $b_{k\ell}$ = number samples being clustered into class k, but actually belonging to class ℓ (by human expertise). The clustering accuracy is $Q = \sum_k b_{kk}/N = 0.875$, quite reasonable for this difficult problem. To provide an understanding of this result, we perform the PCA connectivity analysis. The cluster connectivity matrix P is shown in Fig.3. Clearly, the five smaller classes have strong within-cluster connectivity; the largest class C_1 has substantial connectivity to other classes (those in off-diagonal elements of P). This explains why in clustering results (first column in contingency table B), C_1 is split into several clusters. Also, one tissue sample in C_5 has large connectivity to C_4 and is thus clustered into C_4 (last column in B).

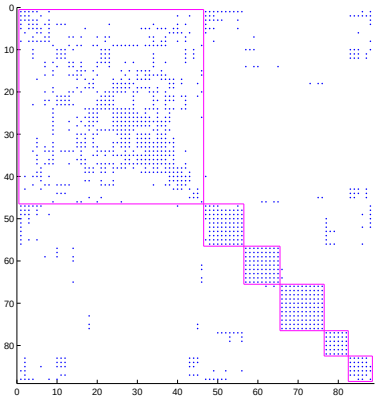

Figure 3. The connectivity matrix for lymphoma. The 6 classes are ordered as $C_1, C_4, C_7, C_9, C_6, C_5$.

Internet Newsgroups

We apply K-means clustering on Internet newsgroup articles. A 20-newsgroup dataset is from www.cs.cmu.edu/afs/cs/project/theo-11/www /naive-bayes.html. Word - document matrix is first constructed. 1000 words are selected according to the mutual information between words and documents in unsupervised manner. Standard `tf.idf` term weighting is used. Each document is normalized to 1.

We focus on two sets of 2-newsgroup combinations and two sets of 5-newsgroup combinations. These four newsgroup combinations are listed below:

```
      A2:                          B2:
NG1:  alt.atheism            NG18: talk.politics.mideast
NG2:  comp.graphics          NG19: talk.politics.misc
      A5:                          B5:
NG2:  comp.graphics          NG2:  comp.graphics
NG9:  rec.motorcycles        NG3:  comp.os.ms-windows
NG10: rec.sport.baseball     NG8:  rec.autos
NG15: sci.space              NG13: sci.electronics
NG18: talk.politics.mideast  NG19: talk.politics.misc
```

In A2 and A5, clusters overlap at medium level. In B2 and B5, clusters overlap substantially.

To accumulate sufficient statistics, for each newsgroup

Table 2. Clustering accuracy as the PCA dimension is reduced from original 1000.

Dim	A5-B	A5-U	B5-B	B5-U
5	0.81/0.91	0.88/0.86	0.59/0.70	0.64/0.62
6	0.91/0.90	0.87/0.86	0.67/0.72	0.64/0.62
10	0.90/0.90	0.89/0.88	0.74/0.75	0.67/0.71
20	0.89	0.90	0.74	0.72
40	0.86	0.91	0.63	0.68
1000	0.75	0.77	0.56	0.57

combination, we generate 10 datasets, each is a random sample of documents from the newsgroups. The details are the following. For A2 and B2, each cluster has 100 documents randomly sampled from each newsgroup. For A5 and B5, we let cluster sizes vary to resemble more realistic datasets. For balanced case, we sample 100 documents from each newsgroup. For the unbalanced case, we select 200,140,120,100,60 documents from different newsgroups. In this way, we generated a total of 60 datasets on which we perform cluster analysis:

We first assess the lower bounds derived in this paper. For each dataset, we did 20 runs of K-means clustering, each starting from different random starts (randomly selecting data points as initial cluster centroids). We pick the clustering results with the lowest K-means objective function value as the final clustering result. For each dataset, we also compute principal eigenvalues of the kernel matrices of X^TX, Y^TY from the uncentered and centered data matrix (see §1).

Table 1 gives the K-means objective function values and the computed lower bounds. Rows starting with Km are the $J_{\rm K}$ optimal values for each data sample. Rows with P2 and P5 are lower bounds computed from Eq.(20). Rows with L2a, L2b are the lower bounds of the earlier work (Zha et al., 2002). L2a are for original data and L2b are for centered data. The last column is the averaged percentage difference between the bound and the optimal value.

For datasets A2 and B2, the newly derived lower-bounds (rows starting with P2) are consistently closer to the optimal K-means values than previously derived bound (rows starting with L2a or L2b).

Across all 60 random samples the newly derived lower-bound (rows starting with P2 or P5) consistently gives close lower bound of the K-means values (rows starting with Km). For $K = 2$ cases, the lower-bound is about 0.6% within the optimal K-means values. As the number of cluster increase, the lower-bound become less tight, but still within 1.4% of the optimal values.

PCA-reduction and K-means

Next, we apply K-means clustering in the PCA subspace. Here we reduce the data from the original 1000 dimensions to 40, 20, 10, 6, 5 dimensions respectively. The clustering accuracy on 10 random samples of each newsgroup combination and size composition are averaged and the results are listed in Table 2. To see the subtle difference between centering data or not at 10, 6, 5 dimensions; results for original uncentered data are listed at left and the results for centered data are listed at right.

Two observations. (1) From Table 2, it is clear that as dimensions are reduced, the results systematically and significantly improves. For example, for datasets A5-balanced, the cluster accuracy improves from 75% at 1000-dim to 91% at 5-dim. (2) For very small number of dimensions, PCA based on the centered data seem to lead to better results. All these are consitent with previous theoretical analysis.

Discussions

Traditional data reduction perspective derives PCA as the best set of bilinear approximations (SVD of Y). The new results show that principal components are continuous (relaxed) solution of the cluster membership indicators in K-means clustering (Theorems 2.2 and 3.1). These two views (derivations) of PCA are in fact consistent since data clustering also is a form of data reduction. Standard data reduction (SVD) happens in Euclidean space, while clustering is a data reduction to classification space (data points in same cluster are considered belonging to same class while points in different clusters are considered belonging to different classes). This is best explained by the vector quantization widely used in signal processing(Gersho & Gray, 1992) where the high dimensional space of signal feature vectors are divided into Voronoi cells via the K-means algorithm. Signal feature vectors are approximated by the cluster centroids, the code-vectors. That PCA plays crucial roles in both types of data reduction provides a unifying theme in this direction.

Acknowledgement

This work is supported by U.S. Department of Energy, Office of Science, Office of Laboratory Policy and Infrastructure, through an LBNL LDRD, under contract DE-AC03-76SF00098.

References

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell

Table 1. K-means objective function values and theoretical bounds for 6 datasets.

	Datasets: A2										
Km	189.31	189.06	189.40	189.40	189.91	189.93	188.62	189.52	188.90	188.19	—
P2	188.30	188.14	188.57	188.56	189.10	188.89	187.85	188.54	187.91	187.25	0.48%
L2a	187.37	187.19	187.71	187.68	188.27	187.99	186.98	187.53	187.29	186.37	0.94%
L2b	185.09	184.88	185.63	185.33	186.25	185.44	185.00	185.56	184.75	184.02	2.13%
	Datasets: B2										
Km	185.20	187.68	187.31	186.47	187.08	186.12	187.12	187.36	185.51	185.50	—
P2	184.44	186.69	186.05	184.81	186.17	185.29	186.13	185.62	184.73	184.19	0.60%
L2a	183.22	185.51	184.97	183.67	185.02	184.19	184.88	184.50	183.55	183.08	1.22%
L2b	180.04	182.97	182.36	180.71	182.46	181.17	182.38	181.77	180.42	179.90	2.74%
	Datasets: A5 Balanced										
Km	459.68	462.18	461.32	463.50	461.71	462.70	460.11	463.24	463.83	463.54	—
P5	452.71	456.70	454.58	457.61	456.19	456.78	453.19	458.00	457.59	458.10	1.31%
	Datasets: A5 Unbalanced										
Km	575.21	575.89	576.56	578.29	576.10	579.12	579.77	574.57	576.28	573.41	—
P5	568.63	568.90	570.10	571.88	569.51	572.26	573.18	567.98	569.32	566.79	1.16%
	Datasets: B5 Balanced										
Km	464.86	464.00	466.21	463.15	463.58	464.70	464.45	465.57	466.04	463.91	—
P5	458.77	456.87	459.38	458.19	456.28	458.23	458.37	458.38	459.77	458.84	1.36%
	Datasets: B5 Unbalanced										
Km	580.14	581.11	580.76	582.32	578.62	581.22	582.63	578.93	578.27	578.30	—
P5	572.44	572.97	574.60	575.28	571.45	574.04	575.18	571.76	571.16	571.13	1.25%

lymphoma identified by gene expression profiling. *Nature*, *403*, 503–511.

Bradley, P., & Fayyad, U. (1998). Refining initial points for k-means clustering. *Proc. 15th International Conf. on Machine Learning*.

Ding, C., & He, X. (2004). Linearized cluster assignment via spectral ordering. *Proc. Int'l Conf. Machine Learning (ICML2004)*.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification, 2nd ed.* Wiley.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*, 183–187.

Fan, K. (1949). On a theorem of Weyl concerning eigenvalues of linear transformations. *Proc. Natl. Acad. Sci. USA*, *35*, 652–655.

Gersho, A., & Gray, R. (1992). *Vector quantization and signal compression*. Kluwer.

Goldstein, H. (1980). *Classical mechanics*. Addison-Wesley. 2nd edition.

Golub, G., & Van Loan, C. (1996). *Matrix computations, 3rd edition*. Johns Hopkins, Baltimore.

Gordon, A., & Henderson, J. (1977). An algorithm for euclidean sum of squares classification. *Biometrics*, 355–362.

Grim, J., Novovicova, J., Pudil, P., Somol, P., & Ferri, F. (1998). Initialization normal mixtures of densities. *Proc. Int'l Conf. Pattern Recognition (ICPR 1998)*.

Hartigan, J., & Wang, M. (1979). A K-means clustering algorithm. *Applied Statistics*, *28*, 100–108.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *Elements of statistical learning*. Springer Verlag.

Jain, A., & Dubes, R. (1988). *Algorithms for clustering data*. Prentice Hall.

Jolliffe, I. (2002). *Principal component analysis*. Springer. 2nd edition.

Lloyd, S. (1957). Least squares quantization in pcm. *Bell Telephone Laboratories Paper, Marray Hill*.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proc. 5th Berkeley Symposium*, 281–297.

Moore, A. (1998). Very fast em-based mixture model clustering using multiresolution kd-trees. *Proc. Neural Info. Processing Systems (NIPS 1998)*.

Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Proc. Neural Info. Processing Systems (NIPS 2001)*.

Schölkopf, B., Smola, A., & Müller, K. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, *10*, 1299–1319.

Wallace, R. (1989). Finding natural clusters through entropy minimization. *Ph.D Thesis. Carnegie-Mellon Uiversity, CS Dept.*

Zha, H., Ding, C., Gu, M., He, X., & Simon, H. (2002). Spectral relaxation for K-means clustering. *Advances in Neural Information Processing Systems 14 (NIPS'01)*, 1057–1064.

Zhang, R., & Rudnicky, A. (2002). A large scale clustering scheme for K-means. *16th Int'l Conf. Pattern Recognition (ICPR'02)*.