

# CUR matrix decompositions for improved data analysis

Michael W. Mahoney<sup>a,1</sup> and Petros Drineas<sup>b</sup>

<sup>a</sup>Department of Mathematics, Stanford University, Stanford, CA 94305; and <sup>b</sup>Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180

Edited by Jon Kleinberg, Cornell University, Ithaca, NY, and accepted by the Editorial Board November 11, 2008 (received for review April 3, 2008)

Principal components analysis and, more generally, the Singular Value Decomposition are fundamental data analysis tools that express a data matrix in terms of a sequence of orthogonal or uncorrelated vectors of decreasing importance. Unfortunately, being linear combinations of up to all the data points, these vectors are notoriously difficult to interpret in terms of the data and processes generating the data. In this article, we develop CUR matrix decompositions for improved data analysis. CUR decompositions are low-rank matrix decompositions that are explicitly expressed in terms of a small number of actual columns and/or actual rows of the data matrix. Because they are constructed from actual data elements, CUR decompositions are interpretable by practitioners of the field from which the data are drawn (to the extent that the original data are). We present an algorithm that preferentially chooses columns and rows that exhibit high “statistical leverage” and, thus, in a very precise statistical sense, exert a disproportionately large “influence” on the best low-rank fit of the data matrix. By selecting columns and rows in this manner, we obtain improved relative-error and constant-factor approximation guarantees in worst-case analysis, as opposed to the much coarser additive-error guarantees of prior work. In addition, since the construction involves computing quantities with a natural and widely studied statistical interpretation, we can leverage ideas from diagnostic regression analysis to employ these matrix decompositions for exploratory data analysis.

randomized algorithms | singular value decomposition | principal components analysis | interpretation | statistical leverage

Modern datasets are often represented by large matrices since an  $m \times n$  real-valued matrix  $A$  provides a natural structure for encoding information about  $m$  objects, each of which is described by  $n$  features. Examples of such objects include documents, genomes, stocks, hyperspectral images, and web groups. Examples of the corresponding features are terms, environmental conditions, temporal resolution, frequency resolution, and individual web users. In many cases, an important step in the analysis of such data is to construct a compressed representation of  $A$  that may be easier to analyze and interpret in light of a corpus of field-specific knowledge. The most common such representation is obtained by truncating the Singular Value Decomposition (SVD) at some number  $k \ll \min\{m, n\}$  terms. For example, Principal Components Analysis (PCA) is just this procedure applied to a suitably normalized data correlation matrix.

Recall the SVD of a general matrix  $A \in \mathbb{R}^{m \times n}$ . Given  $A$ , there exist orthogonal matrices  $U = [u^1 u^2 \dots u^m] \in \mathbb{R}^{m \times m}$  and  $V = [v^1 v^2 \dots v^n] \in \mathbb{R}^{n \times n}$ , where  $\{u^i\}_{i=1}^m \in \mathbb{R}^m$  and  $\{v^i\}_{i=1}^n \in \mathbb{R}^n$  are such that

$$U^T A V = \Sigma = \text{diag}(\sigma_1, \dots, \sigma_\rho),$$

where  $\Sigma \in \mathbb{R}^{m \times n}$ ,  $\rho = \min\{m, n\}$ ,  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_\rho \geq 0$ , and  $\text{diag}(\cdot)$  represents a diagonal matrix with the specified elements on the diagonal. Equivalently,  $A = U \Sigma V^T$ . The 3 matrices  $U$ ,  $V$ , and  $\Sigma$  constitute the SVD of  $A$  (1)—the  $\sigma_i$  are the singular values of  $A$  and the vectors  $u^i$  and  $v^i$  are the  $i$ th left and the  $i$ th right singular vectors, respectively—and  $O(\min\{mn^2, m^2n\})$  time suffices to compute them.

The SVD is widely used in data analysis, often via methods such as PCA, in large part because the subspaces spanned by the vectors

(typically obtained after truncating the SVD to some small number  $k$  of terms) provide the best rank- $k$  approximation to the data matrix  $A$ . If  $k \leq r = \text{rank}(A)$  and we define  $A_k = \sum_{i=1}^k \sigma_i u^i v^{iT}$ , then

$$\|A - A_k\|_F^2 = \min_{X \in \mathbb{R}^{m \times n}: \text{rank}(X) \leq k} \|A - X\|_F^2,$$

i.e., the distance, as measured by the Frobenius norm  $\|\cdot\|_F$ , where  $\|A\|_F^2 = \sum_{ij} A_{ij}^2$ , between  $A$  and any rank  $k$  approximation to  $A$ , is minimized by  $A_k$  (1).

Although the truncated SVD is widely used, the vectors  $u^i$  and  $v^i$  themselves may lack any meaning in terms of the field from which the data are drawn. For example, the eigenvector

$$[(1/2)\text{age} - (1/\sqrt{2})\text{height} + (1/2)\text{income}],$$

being one of the significant uncorrelated “factors” or “features” from a dataset of people’s features, is not particularly informative or meaningful. This fact should not be surprising. After all, the singular vectors are mathematical abstractions that can be calculated for any data matrix. They are not “things” with a “physical” reality.

Nevertheless, data analysts often fall prey to a temptation for reification, i.e., for assigning a physical meaning or interpretation to all large singular components. In certain special cases, e.g., a dataset consisting of points drawn from a multivariate normal distribution on the plane, as in Fig. 1A, the principal components may be interpreted in terms of, e.g., the directions of the axes of the ellipsoid from which the data are drawn. In most cases, however, e.g., when the data are drawn from the union of 2 normals as in Fig. 1B, such reification is not valid. In this and other examples it would be difficult to interpret these directions meaningfully in terms of processes generating the data. Although reification is certainly justified in some cases, such an interpretative claim cannot arise from the mathematics alone, but instead requires an intimate knowledge of the field from which the data are drawn (2).

To understand better the reification issues in modern biological applications, consider a synthetic dataset introduced by Wall *et al.* (3) to model oscillatory and exponentially decaying patterns of gene expression from Cho *et al.* (4). The data matrix consists of 14 expression level assays (columns of  $A$ ) and 2,000 genes (rows of  $A$ ), corresponding to a  $2,000 \times 14$  matrix  $A$ . Genes have 1 of 3 types of transcriptional response: noise (1,600 genes), noisy sine pattern (200 genes), and noisy exponential pattern (200 genes). Fig. 1C and D present the “biological” data, i.e., overlays of 5 noisy sine wave genes and five noisy exponential genes, respectively; Fig. 1E presents first and second singular vectors of the data matrix, along with the original sine pattern and exponential pattern that generated the data; and Fig. 1F shows that the data cluster well in the space spanned by the top 2 singular vectors, which in this case

Author contributions: M.W.M. and P.D. designed research, performed research, contributed new reagents/analytical tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

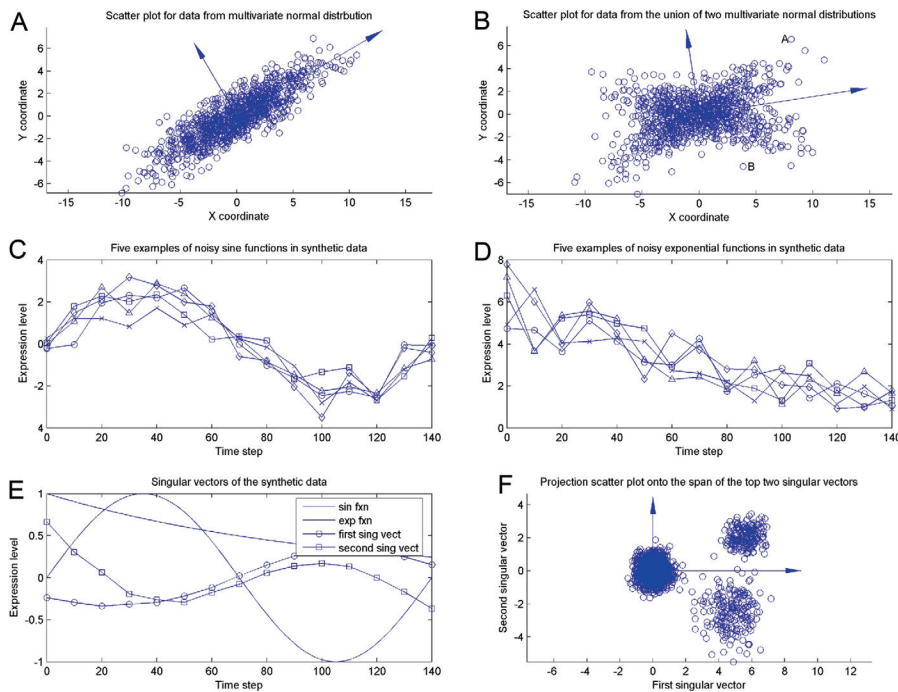
This article is a PNAS Direct Submission.

J.K. is a guest editor invited by the Editorial Board.

<sup>1</sup>To whom correspondence should be addressed. E-mail: mmahoney@cs.stanford.edu.

This article contains supporting information online at [www.pnas.org/cgi/content/full/0803205105/DCSupplemental](http://www.pnas.org/cgi/content/full/0803205105/DCSupplemental).

© 2009 by The National Academy of Sciences of the USA



$$\|A - P_C A\|_F \leq \|A - A_k\|_F + \epsilon \|A\|_F \quad [1]$$

can be obtained, with high probability. Here  $P_C A$  denotes the projection of  $A$  on the subspace spanned by the columns of  $C$ . Subsequently, Drineas, Kannan, and Mahoney (12) constructed an additive-error CUR matrix decomposition by choosing columns and rows simultaneously.\* In 2 passes over the matrix  $A$ , they randomly construct a matrix  $C$  of columns, a matrix  $R$  of rows, and a matrix  $U$  such that

$$\|A - CUR\|_F \leq \|A - A_k\|_F + \epsilon \|A\|_F, \quad [2]$$

with high probability.

The additive-error algorithms of refs. 11 and 12 were motivated by resource-constrained computational environments, e.g., where the input matrices are extremely large or where only a very small fraction of the data is actually available, and in those applications they are appropriate. For example, this additive-error CUR matrix decomposition has been successfully applied to applications such as hyperspectral image analysis, recommendation system analysis, and DNA SNP analysis (13, 14). These additive-error matrix decompositions are, however, quite coarse in the worse case. Moreover, the insights provided by their sampling probabilities into the data are limited—the probabilities are often uniform due to data preprocessing, or they may correspond, e.g., simply to the degree of a node if the data matrix is derived from a graph.

### Statistical Leverage and Improved Matrix Decompositions

To construct  $C$  (similarly  $R$ ), we will compute an “importance score” for each column of  $A$ , and we will randomly sample a small number of columns from  $A$  by using that score as an importance sampling probability distribution. This importance score (see Eq. 3 below) depends on the matrix  $A$ , and it has a natural interpretation as capturing the “statistical leverage” or “influence” of a given column on the best low-rank fit of the data matrix. By preferentially choosing columns that exert a disproportionately large influence on the best low-rank fit (as opposed to procedures that sample columns that have larger Euclidean norm, or empirical variance, as in prior work), we will ensure that  $CUR$  is nearly as good as  $A_k$  at capturing the dominant part of the spectrum of  $A$ . In addition, by choosing “high statistical-leverage” or “highly influential” columns, we can leverage ideas from diagnostic regression analysis to apply CUR matrix decompositions as a tool for exploratory data analysis.

To motivate our choice of importance sampling scores, recall that we can express the  $j$ th column of  $A$  (denoted by  $A^j$ ) exactly as

$$A^j = \sum_{\xi=1}^r (\sigma_{\xi} u^{\xi}) v_j^{\xi},$$

where  $r = \text{rank}(A)$  and where  $v_j^{\xi}$  is the  $j$ th coordinate of the  $\xi$ th right singular vector. That is, the  $j$ th column of  $A$  is a linear combination of all the left singular vectors and singular values, and the elements of the  $j$ th row of  $V$  are the coefficients. Thus, we can approximate  $A^j$  as a linear combination of the top  $k$  left singular vectors and corresponding singular values as

$$A^j \approx \sum_{\xi=1}^k (\sigma_{\xi} u^{\xi}) v_j^{\xi}.$$

\*The algorithms of refs. 11 and 12 provide worst-case additive-error guarantees since the “additional error” in Eqs. 1 and 2 is  $\epsilon \|A\|_F$ , which is independent of the “base error” of  $\|A - A_k\|_F$ . This should be contrasted with a relative-error guarantee of the form Eq. 4, in which the “additional error” is  $\epsilon \|A - A_k\|_F$ , for  $\epsilon > 0$  arbitrarily small, and thus the total error is bounded by  $(1 + \epsilon) \|A - A_k\|_F$ . It should also be contrasted with a constant-factor guarantee of the form Eq. 5, in which the additional error is  $\gamma \|A - A_k\|_F$ , for some constant  $\gamma$ .

Since we seek columns of  $A$  that are simultaneously correlated with the span of all top  $k$  right singular vectors, we then compute the *normalized statistical leverage scores*:

$$\pi_j = \frac{1}{k} \sum_{\xi=1}^k (v_j^{\xi})^2, \quad [3]$$

for all  $j = 1, \dots, n$ . With this normalization, it is straightforward to show that  $\pi_j \geq 0$  and that  $\sum_{j=1}^n \pi_j = 1$ , and thus that these scores form a probability distribution over the  $n$  columns.

Our main algorithm for choosing columns from a matrix—we will call it COLUMNSELECT—takes as input any  $m \times n$  matrix  $A$ , a rank parameter  $k$ , and an error parameter  $\epsilon$ , and then performs the following steps:

1. Compute  $v^1, \dots, v^k$  (the top  $k$  right singular vectors of  $A$ ) and the normalized statistical leverage scores of Eq. 3.
2. Keep the  $j$ th column of  $A$  with probability  $p_j = \min\{1, c\pi_j\}$ , for all  $j \in \{1, \dots, n\}$ , where  $c = O(k \log k / \epsilon^2)$ .
3. Return the matrix  $C$  consisting of the selected columns of  $A$ .

With this procedure, the matrix  $C$  contains  $c'$  columns, where  $c' \leq c$  in expectation and where  $c'$  is tightly concentrated around its expectation. The computation of the column leverage scores uses the top  $k$  right singular vectors of  $A$ , and this computation is the bottleneck in the running time of COLUMNSELECT. It can be performed in time linear in the number of nonzero elements of the matrix  $A$  times a low-degree polynomial in the rank parameter  $k$  (1). We have proven that, with probability at least 99%, this choice of columns satisfies

$$\|A - P_C A\|_F \leq (1 + \epsilon/2) \|A - A_k\|_F, \quad [4]$$

where  $P_C$  denotes a projection matrix onto the column space of  $C$ .<sup>†</sup> See ref. 15 for the proof of Eq. 4, which depends crucially on the use of Eq. 3. In some applications, this restricted CUR decomposition,  $A \approx P_C A = CX$ , where  $X = C^+ A$ , is of interest.

In other applications, one wants such a CUR matrix decomposition in terms of both columns and rows simultaneously. Our main algorithm computing a CUR matrix decomposition—we will call it ALGORITHM CUR—is illustrated in supporting information (SI) Appendix, Fig. S0. This algorithm takes as input any  $m \times n$  matrix  $A$ , a rank parameter  $k$ , and an error parameter  $\epsilon$ , and then it performs the following steps:

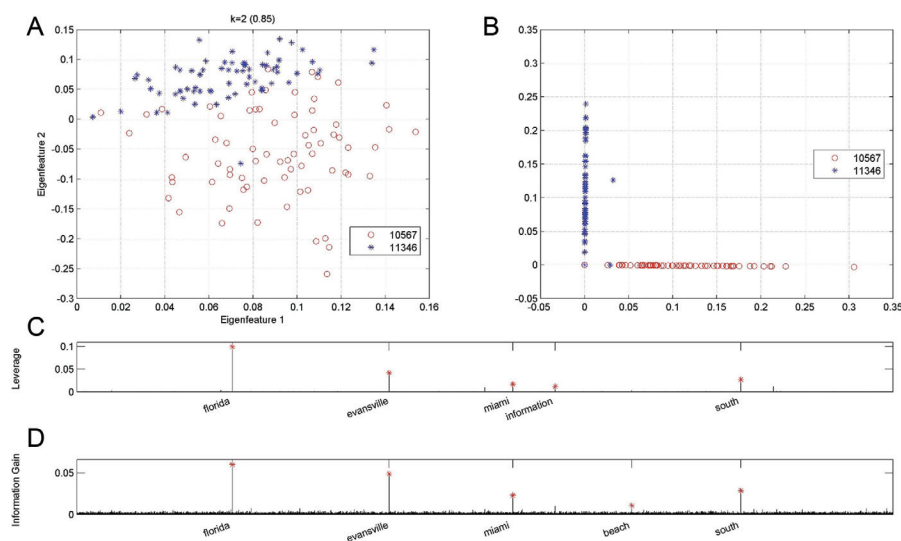
1. Run COLUMNSELECT on  $A$  with  $c = O(k \log k / \epsilon^2)$  to choose columns of  $A$  and construct the matrix  $C$ .
2. Run COLUMNSELECT on  $A^T$  with  $r = O(k \log k / \epsilon^2)$  to choose rows of  $A$  (columns of  $A^T$ ) and construct the matrix  $R$ .
3. Define the matrix  $U$  as  $U = C^+ A R^+$ , where  $X^+$  denotes a Moore–Penrose generalized inverse of the matrix  $X$  (17).

As with our algorithm for selecting columns, the running time of ALGORITHM CUR is dominated by computation of the column and row leverage scores. For this choice of  $C$ ,  $U$ , and  $R$ , we will prove that

$$\|A - CUR\|_F \leq (2 + \epsilon) \|A - A_k\|_F, \quad [5]$$

<sup>†</sup>The randomness and failure probability in COLUMNSELECT and ALGORITHM CUR are over the choices made by the algorithm and not over the input data matrix. The quality of approximation bound Eq. 4 holds for any input matrix  $A$ , regardless of how  $A$  is constructed; its proof relies on matrix perturbation theory (15, 16). The arbitrarily chosen failure probability can be set to any  $\delta > 0$  by repeating the algorithm  $O(\log(1/\delta))$  times and taking the best of the results.





**Fig. 2.** Application of ALGORITHM CUR to Internet term-document data. The matrix consists of 139 documents from TechTC on 2 topics: US:Indiana:Evansville (id:10567) and US:Florida (id:11346). (A) A projection of the documents on the 2D space spanned by the top 2 eigenterms. (B) A projection of the documents on the best 2D approximation to the space spanned by the 5 actual terms with the highest statistical leverage scores of Eq. 3. (C) The statistical leverage scores of the 15,170 terms, with the top 5 terms highlighted. (D) The information gain statistic of the 15,170 terms, with the top 5 terms highlighted.

with probability at least 98%.<sup>‡</sup> First, since  $U = C^+AR^+$ , it immediately follows that

$$\|A - CUR\|_F = \|A - CC^+AR^+\|_F.$$

Adding and subtracting  $CC^+A$  and applying the triangle inequality for the Frobenius norm, we get

$$\begin{aligned} \|A - CUR\|_F &\leq \|A - CC^+A\|_F + \|CC^+A - CC^+AR^+\|_F \\ &\leq \|A - CC^+A\|_F + \|A - AR^+\|_F \\ &= \|A - P_C A\|_F + \|A - AP_R\|_F. \end{aligned} \quad [6]$$

Inequality (6) follows since  $CC^+$  is a projection matrix and thus does not increase the Frobenius norm, and the last equality follows since  $P_C = CC^+$  and similarly  $P_R = R^+R$ . Since ALGORITHM CUR chooses columns and rows by calling COLUMNSELECT on  $A$  and  $A^T$ , respectively, Eq. 5 follows by 2 applications of Eq. 4. Note that  $r = c$  for ALGORITHM CUR, and also that the failure probability for ALGORITHM CUR is at most twice the failure probability of COLUMNSELECT, since the latter algorithm may fail when applied to columns or when applied to rows, independently.

Although one might like to fix a rank parameter  $k$  and choose  $k$  columns and/or rows deterministically according to some criterion—e.g., such as to define a parallelepiped of maximal volume over all  $\binom{n}{k}$  such parallelepipeds, or to span a subspace that “captures” a maximal amount of variance from  $A$  over all  $\binom{n}{k}$  such subspaces—most such criteria would lead to intractable combinatorial optimization problems (9, 10, 18). Thus, ALGORITHM CUR takes advantage of oversampling (choosing slightly more than  $k$  columns) and randomness as computational resources to obtain its strong provable approximation guarantees.

Note that the quantities in Eq. 3 are, up to scaling, equal to the diagonal elements of the so-called “hat matrix,” i.e., the projection matrix onto the span of the top  $k$  right singular vectors of  $A$  (19, 20). As such, they have a natural statistical interpretation as a “leverage score” or “influence score” associated with each of the data points (19–21). In particular,  $\pi_j$  quantifies the amount of leverage or influence exerted by the  $j$ th column of  $A$  on its optimal low-rank approximation. Furthermore, these quantities have been widely used for outlier identification in diagnostic regression analysis (22, 23). Thus, using these scores to select columns not only is crucial for our improved worst-case bounds but also aids in exploratory data analysis.

<sup>‡</sup>This can be improved to  $1 + \epsilon$  by using a somewhat more complicated algorithm (15).

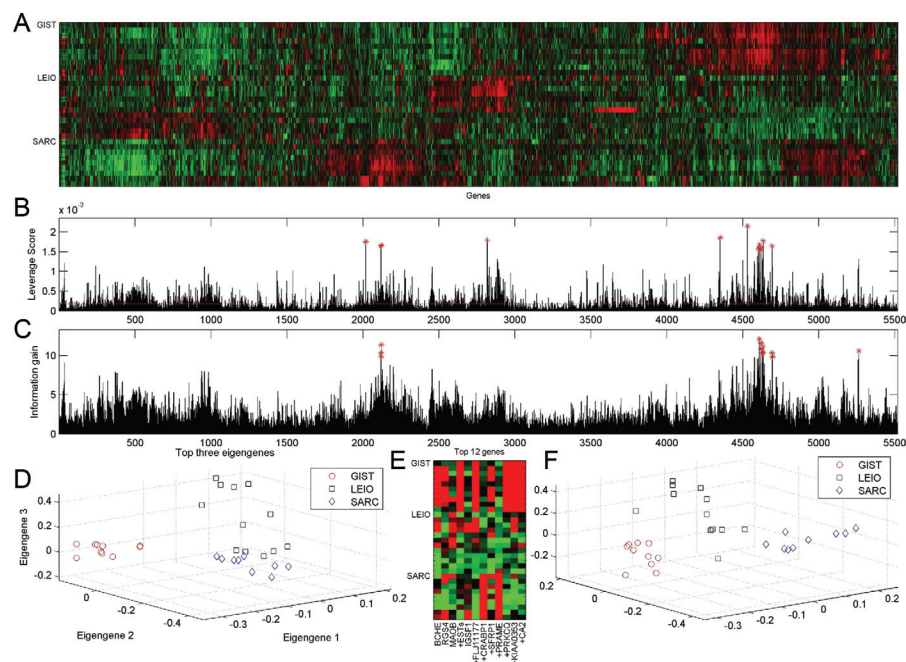
## Diagnostic Data Analysis Applications

Implementation of ALGORITHM CUR is straightforward. We have applied this algorithm to data analysis problems in several application domains: Internet term-document data analysis (see Fig. 2 and SI Appendix, Figs. S2–S5); genetics (see Fig. 3); and social science (see SI Appendix, Fig. S1). In practice, we typically only need to sample a number of columns and/or rows that is a small constant, e.g., between 2 and 4, times the input rank parameter  $k$ . In addition, not only can we perform common data analysis tasks (such as clustering and classification) for which the basis provided by truncating the SVD is often used, but we can also use the normalized leverage scores to explore the data and identify whether there are any disproportionately “important” data elements. (Note that since CUR decompositions are low-rank approximations that use information in the top  $k$  singular subspaces, their domain of applicability is not expected to be broader than that of the SVD.)

Internet term-document data are a common application of SVD-based techniques, often via latent semantic analysis (24, 25). The Open Directory Project (ODP) (26) is a multilingual open content directory of World Wide Web links. TechTC (Technion Repository of Text Categorization Datasets) is a publicly available benchmark set of term-document matrices from ODP with varying categorization difficulty (27) (Table 1). Each matrix of the TechTC dataset encodes information from  $\approx 150$  documents from 2 different ODP categories. To illustrate our method, we focused on 4 datasets such that the documents clustered well into 2 classes when projected in a low-dimensional space spanned by the top few left singular vectors.<sup>§</sup> (See Table 1 for a description of the data and also Fig. 2 and SI Appendix, Figs. S3–S5.)

For example, consider the collection of 139 documents (each described with respect to 15,170 terms) on 2 topics: US:Florida (id:11346) and US:Indiana:Evansville (id:10567). The first topic has 71 documents, and the second has 68; the topics names are descriptive; and the sparsity of the associated document-term matrix is 2.1%. Projecting the documents on the top 2 eigenterms and then running  $k$ -means clustering on the projected data leads to a clustering which has a Pearson correlation coefficient of 0.85 with the (provided) ground truth, is illustrated in Fig. 24.

<sup>§</sup>As is common with term-document data, the TechTC matrices are very sparse, and they are not numerically low-rank. For example, the top 2.5% and 5% of the nonzero singular values of these matrices capture (on average) 5.5% and 12.5% of the Frobenius norm, respectively. For data sets in which a low-dimensional space provided by the SVD failed to capture the category separation, CUR matrix decompositions performed correspondingly poorly.



**Fig. 3.** Application of ALGORITHM<sub>CUR</sub> to cancer microarray data: gene expression patterns of patients with soft tissue tumors analyzed with cDNA microarrays. (A) Raster plot of the data. (B) Statistical leverage scores for each gene. Red stars indicate the 12 genes with the highest leverage scores, and the red dashed line indicates the uniform leverage scores. (C) Information gain for each gene, with the highest scoring genes marked. (D) A projection of the 31 patients in the subspace spanned by the top 3 eigengenes of the full data matrix. (E) Raster plot of the 31 patients with respect to the top 12 genes from B. A (+) sign in front of the name of a gene indicates that this gene was also selected by the (supervised) information gain metric. (F) A 3D plot of the 31 patients in the span of the genes of E.

This class separation implies that the documents are semantically well-represented by low-rank approximation via the SVD, even though they are not numerically low rank. The singular vectors, however, are dense; they contain negative entries; and they are not easily interpretable in terms of natural languages or the ODP hierarchy.

Not only does a CUR matrix decomposition capture the Frobenius norm of the matrix (data not shown), but it can be used to cluster the documents. Fig. 2B shows the 139 documents projected on the best rank 2 approximation to the subspace spanned by the top five “highest-leverage” terms. (Two was chosen as the dimensionality of the low-dimensional space since the documents belong to one of 2 categories—the slowly decaying spectrum provides no guidance in this case. Five was chosen as the number of selected columns based on an analysis of the leverage scores, as described below, and other choices yielded worse results.) In this case, the class separation is quite pronounced—the Pearson correlation coefficient is 0.94, which is improved since CUR provides a low-dimensional space that respects the sparsity in the data. Of course, the data are much more axis-aligned in this low-dimensional space, largely because this space is the span of a small number of actual/interpretable terms, each of which is extremely sparse.

As an example of how we can leverage ideas from diagnostic regression analysis to explore the data, consider Fig. 2C.

which shows the statistical leverage scores of all 15,170 terms. The leverage scores of the top 5 terms—florida (.099482), evansville (.042291), south (.026892), miami (.016890), and information (.011792), as seen in Table 1, are orders of magnitude larger than the uniform leverage scores, equal to  $1/n$ , where  $n = 15,170$  here. This, coupled with the obvious relevance of these terms to the task at hand, *suggests* that these 5 terms are particularly important or influential in this low-dimensional clustering/classification problem. Further evidence supporting this intuition follows from Fig. 2D, which shows the information gain (IG) statistic for each of the 15,170 terms, again with the top 5 terms highlighted. (Recall that the IG for the  $i$ th term is defined as  $IG_i = |f_{i,1} - f_{i,2}|$ , where  $f_{i,1}$  is the frequency of the  $i$ th term within the documents of the first category, and  $f_{i,2}$  is the frequency of the  $i$ th term within the documents of the second category. In particular, note that it is a supervised metric, i.e., the topic of each document is known prior to computing it.)

The truncated SVD has also been applied to gene expression data (3, 5, 28, 29) in systems biology applications (where one wants to understand the relationship between actual genes) and in clinical applications (where one wants to classify tissue samples from actual patients with and without a disease). See Fig. 3A for a raster plot of a typical such dataset (30), consisting of  $m = 31$  patients with 3 different cancer types [gastrointestinal stromal tumor (GIST), leiomyosarcoma (LEIO), and synovial sarcoma

Table 1. Four TechTC matrices that cluster well into the correct 2 topics by using  $k$ -means in the best low-dimensional space

Categories (numeric ID, description)	no. docs $\times$ no. terms	$k$	PCC	1/(no. terms)	High-leverage terms
10567, US: Indiana: Evansville 11346, US: Florida	139 $\times$ 15170	2	.85	.000066	<b>florida</b> (.099482), <b>evansville</b> (.042291), <b>south</b> (.026892), <b>miami</b> (.016890), <b>information</b> (.011792)
11346, US: Florida 22294, Canada: British Columbia: Nanaimo	125 $\times$ 14392	2	.97	.000069	<b>florida</b> (.097158), <b>nanaimo</b> (.085653), <b>south</b> (.026414), <b>miami</b> (.014415), <b>contact</b> (.007828)
20186, US: Texas: Dallas 22294, Canada: British Columbia: Nanaimo	130 $\times$ 12708	2	.90	.000079	<b>dallas</b> (.079332), <b>nanaimo</b> (.071752), <b>information</b> (.007878), <b>texas</b> (.007788), <b>contact</b> (.007616)
22294, Canada: British Columbia: Nanaimo 25575, Asia: Taiwan: Business and Economy	127 $\times$ 10012	2	.88	.000100	<b>nanaimo</b> (.054242), <b>taiwan</b> (.019860), <b>megahome</b> (.004304), <b>contact</b> (.004113), <b>distiller</b> (.003906)

The first 3 columns indicate the 2 topics of the documents in each matrix and its dimensions (each topic has a numeric ID, as well as a short description). The fourth column indicates our choice for the number  $k$  of significant principal components in the data. The fifth column indicates the Pearson correlation coefficient (PCC) between the ground truth and the 2 clusters returned by applying  $k$ -means clustering on the projected data. The sixth column indicates the uniform leverage scores for each term, i.e., each term is assigned the same score. The last column indicates the 5 terms with the highest leverage scores for each matrix, as well as their corresponding scores in parentheses. (Notice that the scores are many orders of magnitude larger than the uniform score.)



# Supplementary Material: CUR matrix decompositions for improved data analysis

Michael W. Mahoney

Petros Drineas

## Illustration of our main algorithm

In Figure 0, we provide an illustration of the action of `ALGORITHMCUR`, our main low-rank CUR matrix decomposition algorithm.

## Additional data applications

To understand better our CUR matrix decomposition as an exploratory data analysis tool, we analyzed social science data: voting data from the “Second Rehnquist” U.S. Supreme Court [7] and U.S. House of Representatives roll call data for the 107<sup>th</sup> Congress [5, 6]. These are two datasets to which SVD-based analysis has been applied. The Supreme Court data consists of a  $468 \times 9$  matrix with entries in  $\{-1, +1\}$  encoding the decisions from about 70% of the cases decided by the 9 members of the U.S. Supreme Court from 1994 to 2003 [7]. The House data consists of a  $\{-1, 0, +1\}$  encoding of the “yea,” not voting, and “nay,” voting records on each of the  $m = 990$  roll call votes for each of the  $n = 444$  representatives (including midterm replacements).

When voting data from the Supreme Court [7]—arguably a data set of such a size and about which too much field specific knowledge is available for random sampling to be appropriate—is plotted on the top two singular vectors, one clearly observes the characteristic “horseshoe” pattern of data that has local ordering information [2]. (See Figures 1(a-c).) If we choose  $k = 2$ , which captures 79.0% of the variance in the data, then the justice with the highest leverage score is Stevens, followed by Thomas and then Scalia, and the pair of judges which captures the most variance in the data are Rehnquist and Ginsburg, who together capture 65.9% of the data variance. If  $k = 1$ , which has been interpreted as ordering the justices along a partisan axis and which captures 57.1% of the variance in the data, then the justices with the highest leverage scores are Kennedy and then O’Connor.

In addition, we can pick 3 actual justices (columns of  $A$ , in this case corresponding to, *e.g.*, Scalia, Kennedy, and Ginsburg) and then 3 actual cases (rows of  $A$ ), such that our *CUR* approximation to  $A$  accurately reconstructs 91.0% of the justices’ opinions. We do so by first computing the low-rank approximation  $A' = CUR$  (or  $A' = A_k$  by truncating the SVD if we are interested in eigenjustices and eigencases) and then rounding each element of  $A'$  to the nearest element in  $\{-1, +1\}$ . For comparison, by keeping just the top 2 eigenjustices and eigencases, 92.8% of the entries of the Supreme Court data matrix are accurately predicted [7].

Congressional roll call data are known to be much more homogenous [5, 6]. (See Figures 1(d) and 1(e).) For example, if  $k = 2$ , which captures 75% of the data variance, then no representative exhibits a particularly high leverage, *e.g.*, none has a leverage score greater than  $2k/n$ . Indeed, if  $k = 2$ , the highest leverage score for any representative is only  $1.38k/n$ . (A value of greater than  $2k/n$  is of interest since it has been suggested as a rule of thumb to identify outliers [3].) In this case, uniform random sampling does nearly as well (both in theory and practice) as nonuniform



sampling based on the leverage scores, and selecting in a greedy manner representatives with the highest leverage scores does quite poorly. For example, the projection of the members of the U.S. House of Representatives onto the span of two representatives chosen by randomly sampling according to the “leverage score” probabilities is very similar to the plot obtained for the projection onto the “best pair” of representatives (DeLauro (D-CT) and Crenshaw(R-FL)). It is also very similar to the plot obtained for *any* two representatives chosen from as the best of 10 trials, with each trial performed by sampling uniformly at random. Applying the  $k$ -means clustering algorithm on the two-dimensional data of this panel, we easily separate the Democrats from the Republicans, with only the four labeled misclassifications. For comparison,  $k$ -means on the full dataset misclassifies two or three representatives.

On the other hand, many other data sets exhibit leverage scores that are extremely nonuniform (from the point of view of classical regression diagnostics). Two were discussed in the main text. In addition, consider the full Enron email data. (See Figure 2.) The data matrix is constructed from the PRIVATE collection [1] using log-entropy term weighing and consists of  $m = 65,033$  messages and a total of  $n = 92,133$  terms. In this case, if, *e.g.*,  $k = 12$ , then 6,029 terms (out of 92,133) have a leverage score greater than  $2k/n$ , 981 terms have a leverage score greater than  $20k/n$ , and 29 terms have a leverage score greater than  $200k/n$ . Consequently, uniform sampling does very poorly, when compared with nonuniform random sampling, but greedily keeping the highest leverage terms—the greatest outliers—does remarkably well at capturing the variance in the data. For example, in Figure 2(b), the rank parameter is fixed, and shown is the best of 10 randomized trials. Results shown are for setting  $k = 12$ , which captures only 13.3% of the variance of the data; similar results are obtained for other values of  $k$ , but note that, *e.g.*,  $k = 60$  and  $k = 120$  still capture only 22.1% and 27.8%, respectively, of the data variance. Interestingly, greedily keeping the terms with the highest leverage scores also does well in applications of more immediate interest to the data analyst. For example, terms with leverage scores greater than, *e.g.*, *ca.*  $10k/n$  are less correlated with the main axis of variance of data and thus tend to be much more interesting and discriminative if one is interested in tasks such as the detection of time evolving, novel, or outlier topics. (See Figure 2(b) for details.) We saw something similar in the main text.

Finally, see Figures 3, 4, and 5 for additional figures illustrating the application of ALGORITHM CUR to other TechTC datasets that cluster well in the low-dimensional space.



## References

- [1] M.W. Berry and M. Browne. Email surveillance using non-negative matrix factorization. *Computational and Mathematical Organization Theory*, 11(3):249–264, 2005.
- [2] P. Diaconis, S. Goel, and S. Holmes. Horseshoes in multidimensional scaling and kernel methods. *To appear in: Annals of Applied Statistics*.
- [3] D.C. Hoaglin and R.E. Welsch. The hat matrix in regression and ANOVA. *The American Statistician*, 32(1):17–22, 1978.
- [4] B. Klimt and Y. Yang. The Enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning*, pages 217–226, 2004.
- [5] K.T. Poole and H. Rosenthal. Patterns of congressional voting. *American Journal of Political Science*, 35:228–278, 1991.
- [6] M.A. Porter, P.J. Mucha, M.E.J. Newman, and C.M. Warmbrand. A network analysis of committees in the U.S. House of Representatives. *Proc. Natl. Acad. Sci. USA*, 102(20):7057–7062, 2005.
- [7] L. Sirovich. A pattern analysis of the second Rehnquist U.S. Supreme Court. *Proc. Natl. Acad. Sci. USA*, 100(13):7432–7437, 2003.

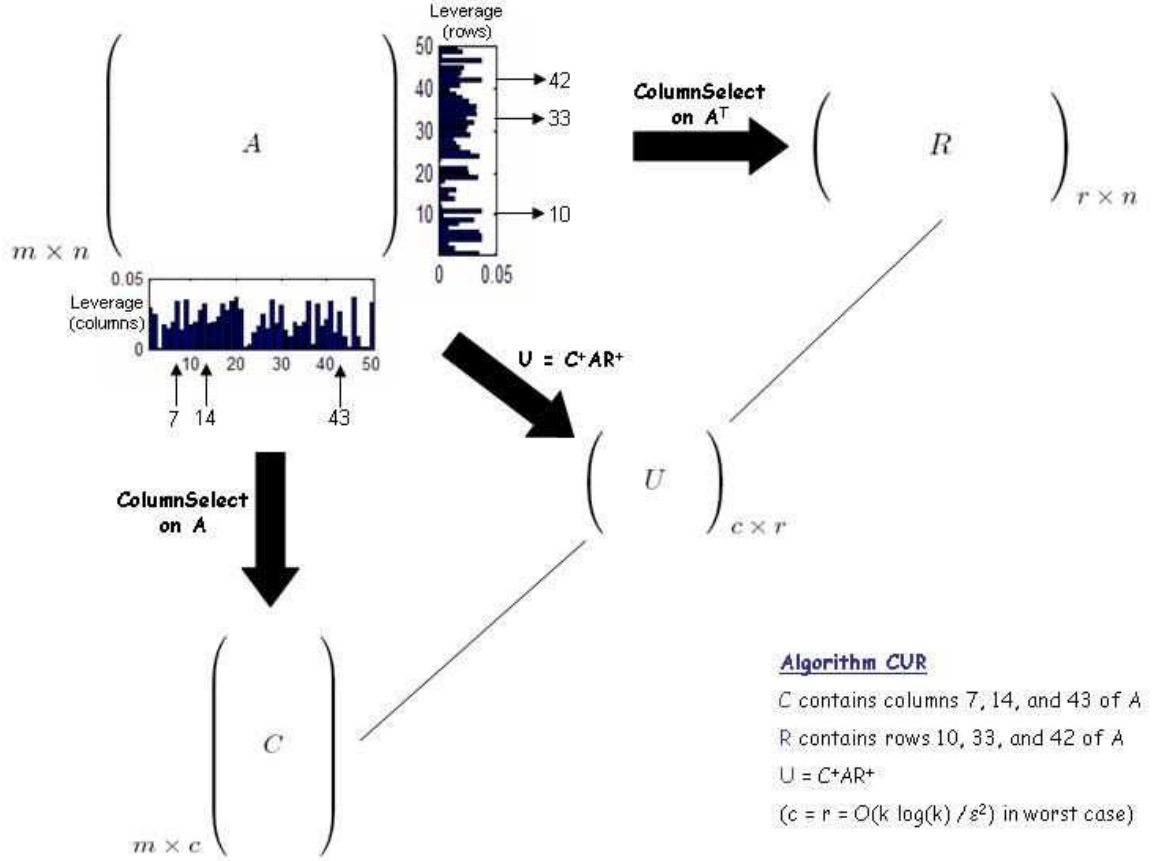


Figure 0: ALGORITHMCUR, our main low-rank CUR matrix decomposition algorithm. It takes as input an  $m \times n$  real-valued matrix  $A$ , a rank parameter  $k$ , and an accuracy parameter  $\epsilon$ . Its output consists of three matrices,  $C$ ,  $U$ , and  $R$  such that the error  $\|A - CUR\|_F$  is at most  $(2 + \epsilon)$  times the error of the best rank  $k$  approximation.

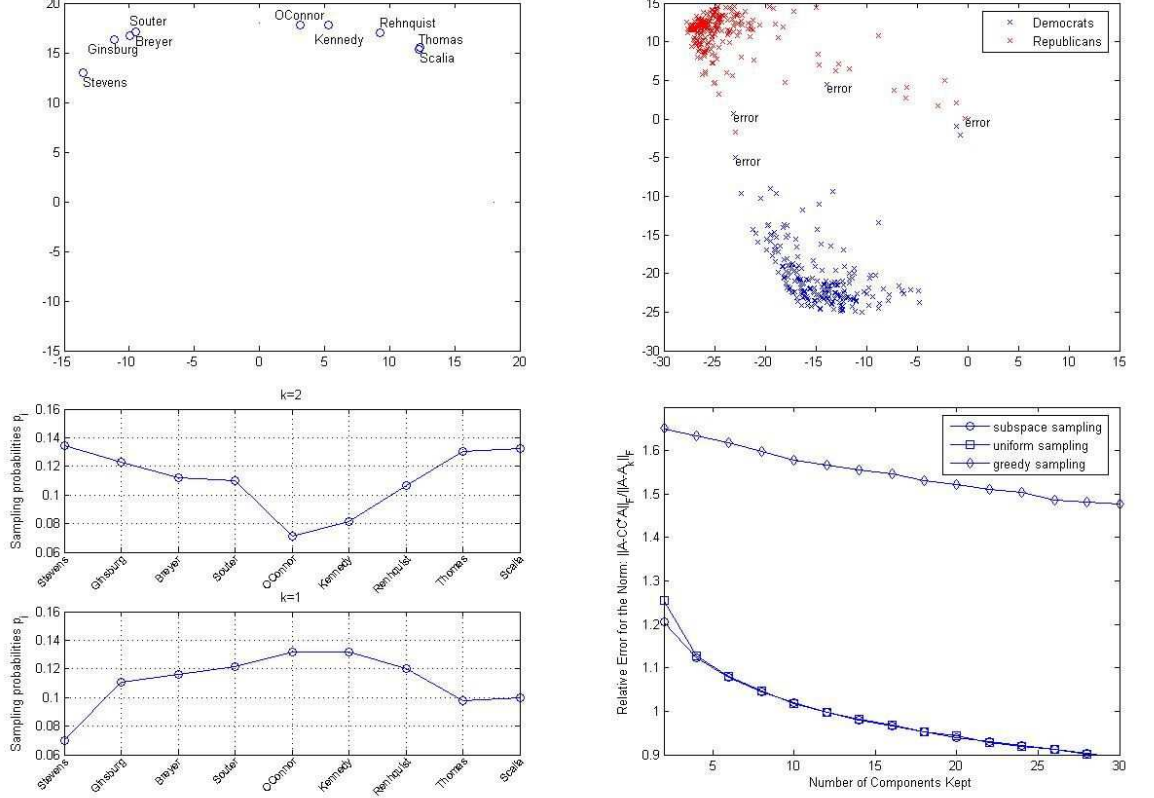


Figure 1: Application to social science data: voting data from the “Second Rehnquist” U.S. Supreme Court [7] and U.S. House of Representatives roll call data for the 107<sup>th</sup> Congress [5, 6]. (A) Projection of the justices onto the space spanned by the top  $k = 2$  singular justices. The  $x$ -axis corresponds to the top eigenjustice and the  $y$ -axis corresponds to the second eigenjustice. (B) and (C) Statistical leverage scores for each of the judges for the choice of  $k = 2$  and  $k = 1$ . (D) Projection of the members of the U.S. House of Representatives onto the span of two representatives chosen by randomly sampling according to the “leverage score” probabilities; shown is the best of 10 independent trials. The  $x, y$ -axes correspond to a coordinate system for the subspace spanned by the chosen two representatives. (E) Frobenius norm error for randomly sampling with the “leverage score” probabilities, compared with uniform random sampling and greedily selecting the representatives with the highest leverage scores. For random sampling, shown is the best of 10 independent trials.

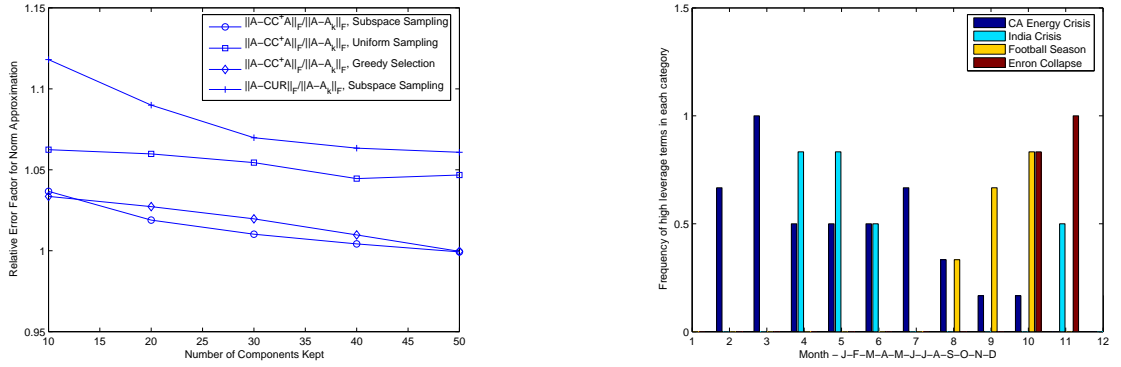


Figure 2: Application to Internet data: term-document data matrix derived from the Enron email corpus [4]. (A) Frobenius norm error for: randomly sampling columns as in ALGORITHM CUR; randomly sampling columns and then rows as in ALGORITHM CUR; uniform random sampling; and greedily selecting the terms with the highest leverage scores. (B) Detection of Outlier Topics: The full PRIVATE collection for 2001 is decomposed into 12 submatrices, one for each month. Then, for each month, the “leverage scores” are used to order terms from highest to lowest. Shown is a bar chart corresponding to the frequency with which terms in that category appear in the top 50 highest leverage terms. For simplicity six terms are used to define each topic: CA Energy Crisis (power, california, electricity, utilities, market, customers); India Crisis (dabhol, dpc, india(n), lender, mseb, maharashtra); Football Season (texas, fantasy, game, ut, orange, longhorns); and Enron Collapse (partnership, fastow, shares, sec, stock, investor).



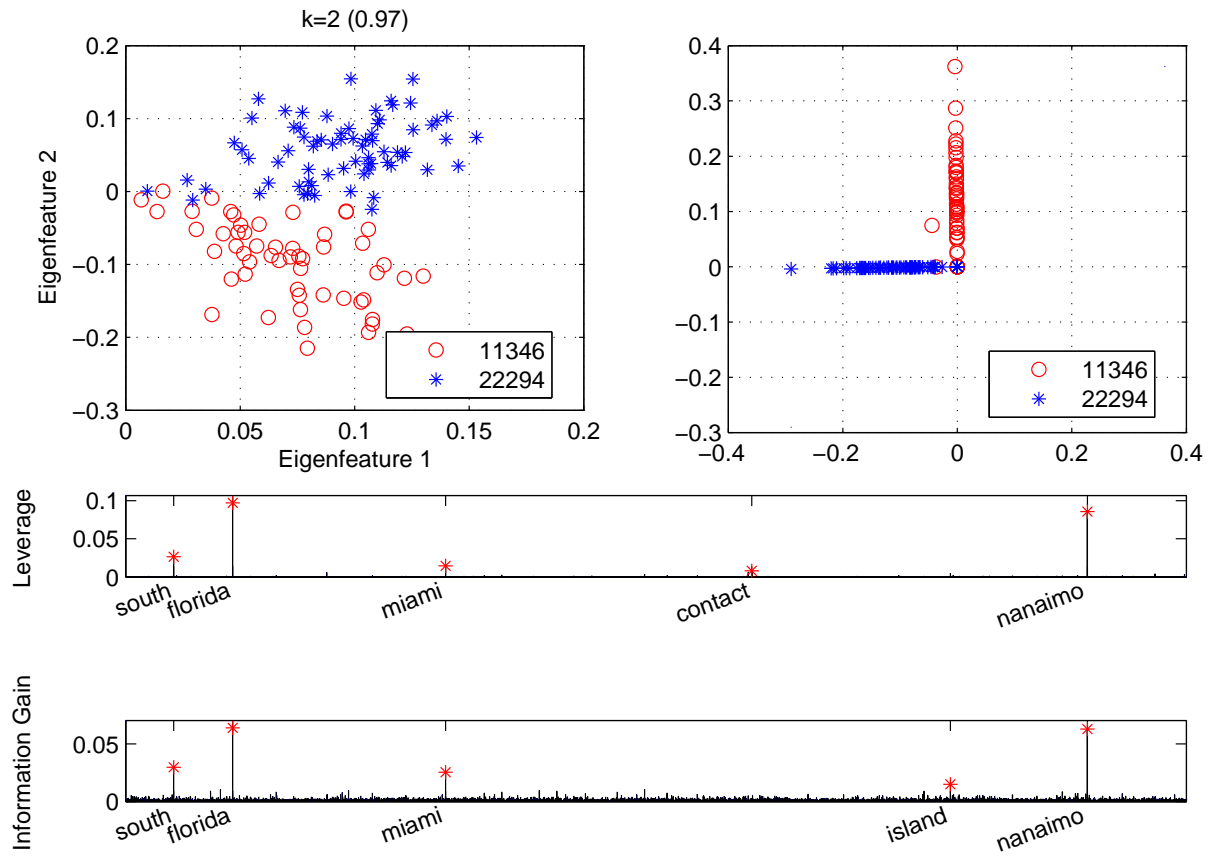


Figure 3: Application of ALGORITHM CUR to a TechTC dataset. The matrix consists of 125 documents from TechTC on two topics: US:Florida (id:11346) and Canada:British Columbia:Nanaimo (id:22294).

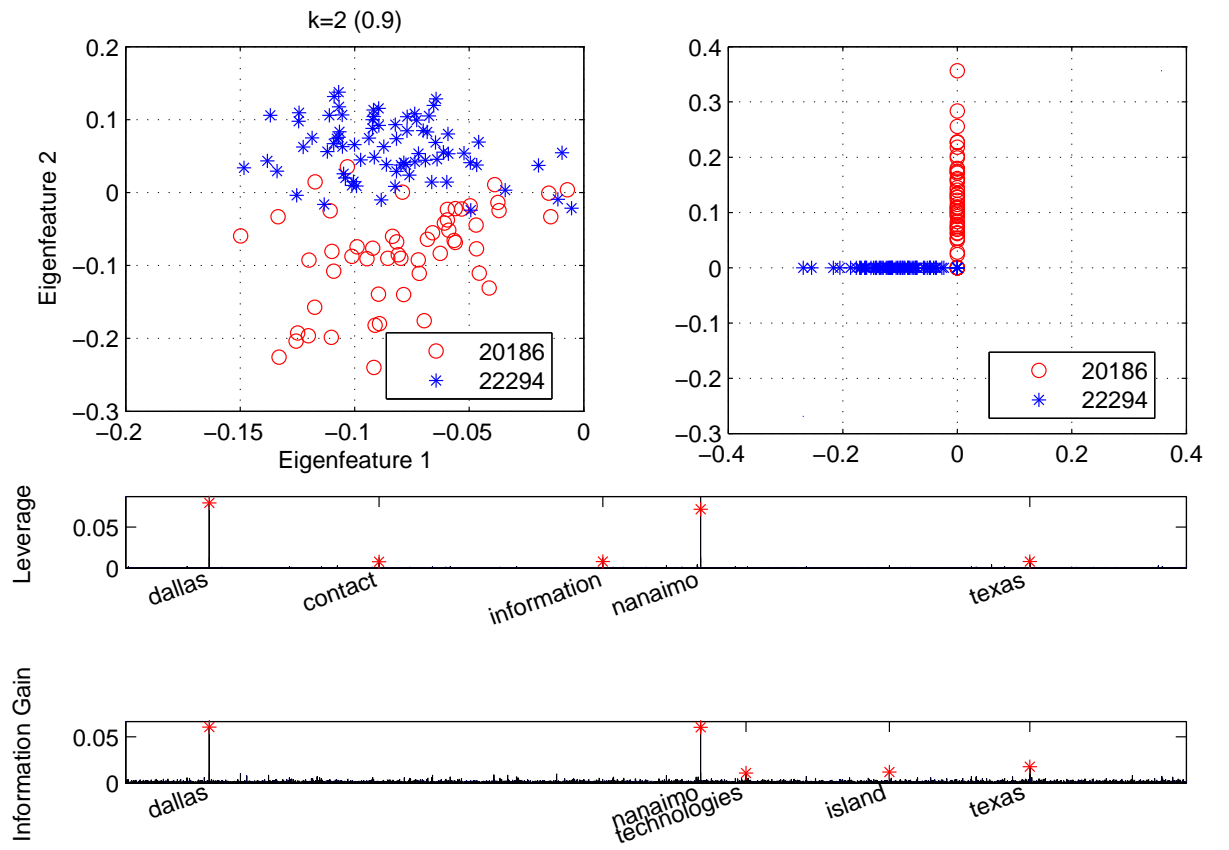


Figure 4: Application of ALGORITHM CUR to a TechTC dataset. The matrix consists of 130 documents from TechTC on two topics: US:Texas:Dallas (id:20186) and Canada:British Columbia:Nanaimo (id:22294).

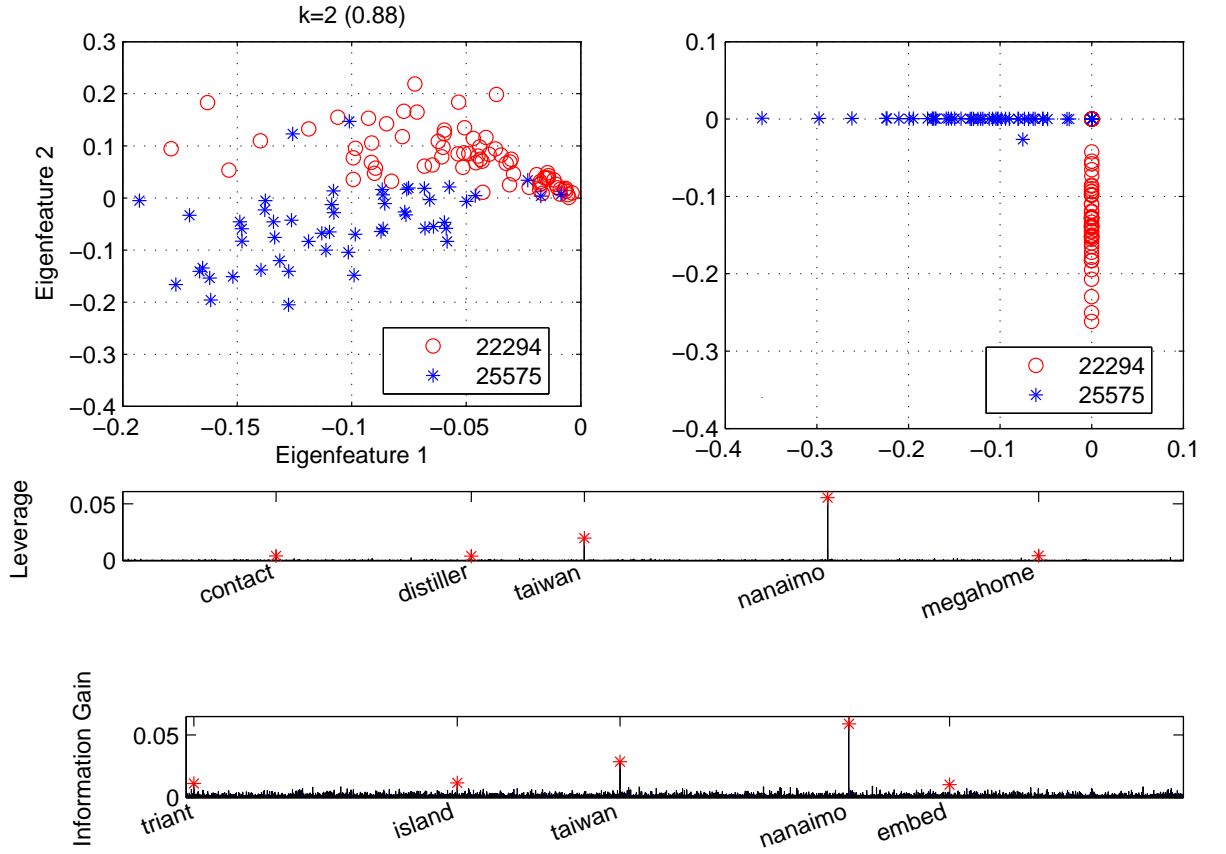


Figure 5: Application of ALGORITHM CUR to a TechTC dataset. The matrix consists of 127 documents from TechTC on two topics: Canada:British Columbia:Nanaimo (id:22294) and Asia:Taiwan:Business and Economy (id:25575).