Subspace sampling and relative-error matrix approximation

Petros Drineas

Rensselaer Polytechnic Institute
Computer Science Department

(joint work with M. W. Mahoney)

For papers, etc.

YAHOO! drineas
The CUR decomposition

\[
\begin{pmatrix}
A \\
\end{pmatrix}_{m \times n} \approx \begin{pmatrix}
C \\
\end{pmatrix} \cdot \begin{pmatrix}
U \\
\end{pmatrix} \cdot \begin{pmatrix}
R \\
\end{pmatrix}
\]

\text{Goal: make (some norm) of } A - \text{CUR small.}

\text{Why? After making two passes over } A, \text{ we can compute provably good } C, U, \text{ and } R \text{ and store them ("sketch") instead of } A: O(m+n) \text{ vs. } O(mn) \text{ RAM space.}

\text{Why? Given a sample consisting of a few columns (C) and a few rows (R) of } A, \text{ we can compute } U \text{ and "reconstruct" } A \text{ as CUR.}

\text{If the sampling probabilities are not "too bad", we get provably good accuracy.}

\text{Why? Given sufficient time, we can find } C, U \text{ and } R \text{ such that } A - \text{CUR is almost optimal.}

\text{This might lead to improved data interpretation.}
Overview

- Background & Motivation
- Relative error $CX$ and $CUR$
- Open problems
Singular Value Decomposition (SVD)

\[
\begin{pmatrix}
A \\
\end{pmatrix}_{m \times n} = 
\begin{pmatrix}
U \\
\end{pmatrix}_{m \times \rho} \cdot 
\begin{pmatrix}
\Sigma \\
\end{pmatrix}_{\rho \times \rho} \cdot 
\begin{pmatrix}
V \\
\end{pmatrix}_{\rho \times n}
\]

\[\tau<\text{rank of } A\]

\[U (V): \text{orthogonal matrix containing the left (right) singular vectors of } A.\]

\[\Sigma: \text{diagonal matrix containing the singular values of } A.\]

Exact computation of the SVD takes \(O(\min\{mn^2, m^2n\})\) time.

The top \(k\) left/right singular vectors/values can be computed faster using Lanczos/Arnoldi methods.
Singular Value Decomposition (SVD)

\[
\begin{pmatrix}
A \\
m \times n
\end{pmatrix}
= 
\begin{pmatrix}
U \\
m \times \rho
\end{pmatrix}
\cdot 
\begin{pmatrix}
\Sigma \\
\rho \times \rho
\end{pmatrix}
\cdot 
\begin{pmatrix}
V \\
\rho \times n
\end{pmatrix}^T
\]

\(\tau<\text{rank of } A\)

\(U (V)\): orthogonal matrix containing the left (right) singular vectors of \(A\).

\(Y\): diagonal matrix containing the singular values of \(A\).

Pseudoinverse of \(A\):

\[
A^+ = V \, Y^{-1} \, U^T
\]
**Rank k approximations** \((A_k)\)

\[
\begin{pmatrix}
A_k \\
\end{pmatrix}_{m \times n} = 
\begin{pmatrix}
U_k \\
\end{pmatrix}_{m \times k} \cdot 
\begin{pmatrix}
Σ_k \\
\end{pmatrix}_{k \times k} \cdot 
\begin{pmatrix}
V_k^T \\
\end{pmatrix}_{k \times n}
\]

\(U_k (V_k)\): orthogonal matrix containing the top k left (right) singular vectors of \(A\).

\(Σ_k\): diagonal matrix containing the top k singular values of \(A\).

\(A_k\) is a matrix of rank \(k\) such that \(\|A - A_k\|_F\) is minimized over all rank \(k\) matrices.

**Definition:**

\[
\|A\|_F^2 = \sum_{i,j} A_{i,j}^2
\]
$U_k$ and $V_k$

$A_k = \left( \begin{array}{c} U_k \ \\ \Sigma_k \ \\ V_k^T \end{array} \right)$

$A_k$: matrix of size $m \times n$.

$U_k$: orthogonal matrix containing the top $k$ left (right) singular vectors of $A$.

$V_k^T$: orthogonal matrix containing the top $k$ right (left) singular vectors of $A$.

$\Sigma_k$: diagonal matrix containing the top $k$ singular values of $A$.

The rows of $V_k^T$ are linear combinations of all rows of $A$.

The columns of $U_k$ are linear combinations of all columns of $A$.

$U_k$ ($V_k$): orthogonal matrix containing the top $k$ left (right) singular vectors of $A$.

$Y_k$: diagonal matrix containing the top $k$ singular values of $A$. 
Potential problems with SVD

Structure in the data is not respected by mathematical operations on the data:
- Reification - maximum variance directions are just that.
- Interpretability - what does a linear combination of 6000 genes mean.
- Sparsity - is destroyed by orthogonalization.
- Non-negativity - is a convex and not linear algebraic notion.

Do there exist “better” low-rank matrix approximations?
- “better” structural properties for certain applications.
- “better” at respecting relevant structure.
- “better” for interpretability and informing intuition.
Exploit **structural properties of CUR** to analyze human genomic data:

We seek subjects and SNPs that capture most of the diversity in the data:

- Singular vectors are useless; linear combinations of humans and/or SNPs make no biological sense.
- **CUR** extracts a low-dimensional representation in terms of subjects and SNPs.

**Human genomic data:** Paschou (Yale U), Mahoney (Yahoo! Research),..., Kidd (Yale U), & D. '06.
Prior work: additive error CUR
(D. & Kannan ’03, D., Mahoney, & Kannan ’05)

Let $A_k$ be the “best” rank $k$ approximation to $A$. Then, after two passes through $A$, we can pick $O(k/\gamma^4)$ rows and $O(k/\gamma^4)$ columns, such that

$$\|A - CUR\|_F \leq \|A - A_k\|_F + \varepsilon \|A\|_F \gg \|A - A_k\|_F$$

Additive error is **prohibitively large** in data analysis applications! This “coarse” CUR does not capture the relevant structure in the data.
Theorem: relative error CUR

(D., Mahoney, & Muthukrishnan ’05, ’06)

For any \( k \), \( O(SVD_k(A)) \) time suffices to construct \( C, U, \) and \( R \) s.t.

\[
\| A - CUR \|_F \leq (1 + \varepsilon) \| A - U_k \Sigma_k V_k^T \|_F = (1 + \varepsilon) \| A - A_k \|_F
\]

holds with probability at least \( 1 - \phi \), by picking

\[
O( k \log k \log(1/\phi) / \gamma^2 ) \text{ columns, and } O( k \log^2 k \log(1/\phi) / \gamma^6 ) \text{ rows.}
\]

\( O(SVD_k(A)) \): time to compute the top \( k \) left/right singular vectors and values of \( A \).
Applications: relative error CUR

Evaluation on:
- Microarray data (yeast), a (roughly) 6200 × 24 matrix. (from O. Alter, UT Austin)
- Genetic marker data, 38 matrices, each (roughly) 60 × 65 (with P. Paschou, Yale U.)
- HapMap SNP data, 4 matrices, each (roughly) 70 × 800 (with P. Paschou, Yale U.)

For (small) \( k \), in \( O(\text{SVD}_k(A)) \) time we can construct \( C, U, \) and \( R \) s.t.

\[
\| A - CUR \|_F \leq (1 + .001) \| A - A_k \|_F
\]

by typically picking at most \((k+5)\) columns and at most \((k+5)\) rows.
**CX matrix decompositions**

Create an approximation to $A$ using columns of $A$

\[ A \approx \begin{pmatrix} C \\ X \end{pmatrix} \]

$c=O(1)$ columns

**Goal:** Provide almost optimal bounds for some norm of $A - CX$.

1. How do we draw the columns of $A$ to include in $C$?
2. How do we construct $X$? One possibility is

\[
\min_{X \in \mathbb{R}^{c \times n}} \|A - CX\|_F = \|A - C \left(C^+ A\right)\|_F
\]
Subspace sampling

\[
\begin{pmatrix}
A_k
\end{pmatrix}_{m \times n} = \begin{pmatrix}
U_k
\end{pmatrix}_{m \times k} \cdot \begin{pmatrix}
\Sigma_k
\end{pmatrix}_{k \times k} \cdot \begin{pmatrix}
V_k^T
\end{pmatrix}_{k \times n}
\]

\text{\(U_k(V_k)\): orthogonal matrix containing the top \(k\) left (right) singular vectors of \(A\).}

\(Y_k\): diagonal matrix containing the top \(k\) singular values of \(A\).
Subspace sampling

\[ V_k \]
\[ n \times k \]

Note: The columns of \( V_k \) are orthonormal vectors, **BUT** the rows of \( V_k \) (notation: \( (V_k)_{(i)} \)) are not orthonormal vectors.

**Subspace sampling** in \( O(\text{SVD}_k(A)) \) time

\[
\forall i = 1, 2, \ldots, n \quad p_i = \frac{\|(V_k)_{(i)}\|^2}{k}
\]
Relative-error CX decomposition

- Compute the probabilities $p_i$;
- For each $i = 1, 2, \ldots, n$, pick the $i$-th column of $A$ with probability $\min\{1, c p_i\}$
- Let $C$ be the matrix containing the sampled columns;

$(C$ has $\cdot c$ columns in expectation$)$

**Theorem:** For any $k$, let $A_k$ be the “best” rank $k$ approximation to $A$.

In $O(\text{SVD}_k(A))$ we can compute $p_i$ such that if $c = O(k \log k / \gamma^2)$ then, with probability at least $1 - \delta$,

$$
\min_{X \in \mathbb{R}^{c \times n}} \|A - CX\|_F = \|A - CC^+A\|_F \\
\leq (1 + \varepsilon)\|A - A_k\|_F
$$
Inside subspace sampling

Let \( C = AS \), where \( S \) is a sampling/rescaling matrix and let the SVD of \( A \) be \( A = U_A \Sigma_A V_A^T \). Then,

\[
A - C \left( C^+ A \right) = A - AS \left( AS \right)^+ A
\]

\[
= A - U_A \Sigma_A V_A^T S \left( U_A \Sigma_A V_A^T S \right)^+ A
\]

\[
\| A - C \left( C^+ A \right) \|_F = \| \Sigma_A - \Sigma_A V_A^T S \left( \Sigma_A V_A^T S \right)^+ \Sigma_A \|_F
\]
Submatrices of orthogonal matrices

Important observation: our subspace sampling probabilities guarantee that $SV_A$ is a full-rank, approx. orthogonal matrix:

$$(SV_A)^T (SV_A)^{1/4} I.$$ 

(Frieze, Kannan, Vempala ‘98, D., Kannan, Mahoney ‘01, 04’, Rudelson, Virshyin ‘05 and even earlier by Bourgain, Kashin, and Tzafriri using uniform sampling.)

This property allows us to completely capture the subspace spanned by the top $k$ right singular vectors of $A$. 
Relative-error CX & low-rank approximations

November 2005: Drineas, Mahoney, and Muthukrishnan
- First relative-error CX matrix factorization algorithm.
- $O(\text{SVD}_k(A))$ time and $O(k^2)$ columns.

January 2006: Har-Peled
- $O(mn k^2 \log k)$ - “linear in mn” time to get $1+\gamma$ approximation.

March 2006: Deshpande and Vempala
- $O(k \log k)$ passes, $O(Mk^2)$ time and $O(k \log k)$ columns.

April 2006: Drineas, Mahoney, and Muthukrishnan
- Improved the DMM November 2005 result to $O(k \log k)$ columns.

April 2006: Sarlos
- Relative-error low-rank approximation in just two passes with $O(k \log k)$ columns, after some preprocessing.
Relative-error CUR decomposition

Create an approximation to $A$, using rows and columns of $A$

$$A \approx \begin{pmatrix} C \end{pmatrix} \cdot \begin{pmatrix} U \end{pmatrix} \cdot \begin{pmatrix} R \end{pmatrix}$$

Goal: Provide very good bounds for some norm of $A - CUR$.

1. How do we draw the columns and rows of $A$ to include in $C$ and $R$?
2. How do we construct $U$?
Step 1: subspace sampling for $C$

Relative-error $CX$ decomposition (given $A$, construct $C$)

- Compute the probabilities $p_i$;
- For each $i = 1, 2, ..., n$, pick the $i$-th column of $A$ with probability $\min\{1, cp_i\}$
- Let $C$ be the matrix containing the sampled columns;

$(C$ has $\cdot c$ columns in expectation$)$
Subspace sampling for $\mathbb{R}$

\[
\begin{pmatrix}
C \\
m \times \tilde{c}
\end{pmatrix}
= \begin{pmatrix}
U_C \\
m \times \rho
\end{pmatrix}
\cdot
\begin{pmatrix}
\Sigma_C \\
\rho \times \rho
\end{pmatrix}
\cdot
\begin{pmatrix}
V_C \\
\rho \times \tilde{c}
\end{pmatrix}^T
\]

$U_C$: orthogonal matrix containing the left singular vectors of $C$.
$t < \text{rank of } C$.

Let $(U_C)_{(i)}$ denote the $i$-th row of $U$. 
Subspace sampling for $\mathbb{R}$

$\begin{pmatrix}
U_C \\
m \times \rho
\end{pmatrix}$

$U_C$ : orthogonal matrix containing the left singular vectors of $C$.

$\rho < \text{rank of } C$.

Let $(U_C)_{(i)}$ denote the $i$-th row of $U$.

Subspace sampling in $O(c^2m)$ time

\[ q_i = \frac{\| (U_C)_{(i)} \|_2^2}{\rho} \quad \forall i = 1, 2, \ldots, m \]
Step 2: constructing U and R

Relative-error CX decomposition (given A, construct C)

• Compute the probabilities $p_i$;
• For each $i = 1, 2, \ldots, n$, pick the $i$-th column of $A$ with probability $\min\{1, cp_i\}$;
• Let $C$ be the matrix containing the sampled columns;

$(C$ has $\cdot c$ columns in expectation$)$

CUR Algorithm (given A and C, return U and R)

• Compute the probabilities $q_i$;
• For each $i = 1, 2, \ldots, m$ pick the $i$-th row of $A$ with probability $\min\{1, rq_i\}$;
• Let $R$ be the matrix containing the sampled rows;
• Let $W$ be the intersection of $C$ and $R$;
• Let $U$ be a (rescaled) pseudo-inverse of $W$;

$(R$ has $\cdot r$ rows in expectation$)$
Overall decomposition

$$
\begin{pmatrix}
A \\
m \times n
\end{pmatrix} \approx
\begin{pmatrix}
C \\
m \times \tilde{c}
\end{pmatrix} \cdot \begin{bmatrix}
D \\
W
\end{bmatrix}^+ \cdot
\begin{pmatrix}
D \\
R \\
\tilde{r} \times n
\end{pmatrix}
$$

- **Columns of $A$**
- **Diagonal rescaling matrix**
- **Rows of $A$**
- **"Intersection" of $C$ and $R$**
- $\tilde{r} \times \tilde{c}$
Analyzing Step 2 of CUR

**CUR Algorithm** (given A and C, return U and R)
- Compute the probabilities $q_i$;
- For each $i = 1, 2, \ldots, m$ pick the $i$-th row of $A$ with probability $\min\{1, rq_i\}$;
- Let $R$ be the matrix containing the sampled rows;
- Let $W$ be the intersection of $C$ and $R$;
- Let $U$ be a (rescaled) pseudo-inverse of $W$;

(R has $r$ rows in expectation)

**Theorem:** Given $C$, in $O(c^2m)$ time, we can compute $q_i$ such that

$$
\left\| A - C \underbrace{\left( DW \right)^+ D R}_{U} \right\|_F \leq (1 + \epsilon) \left\| A - C \left( C^+ A \right) \right\|_F
$$

holds with probability at least $1-\phi$, if $r= O(c \log c / \gamma^2)$ rows.
Putting the two theorems together

**Thm 1:** For any $k$, let $A_k$ be the “best” rank $k$ approximation to $A$.

Then, in $O(\text{SVD}_k(A))$ we can pick (in expectation) $c = O(k \log k / \gamma^2)$ columns of $A$ such that, with probability at least $1 - \phi$,

$$\|A - C \left( C^+ A \right) A \|_F \leq (1 + \varepsilon) \|A - A_k\|_F$$

**Thm 2:** Given $A$ and $C$, in $O(c^2m)$ time, we can pick (in expectation) $r = O(c \log c / \gamma^2)$ rows of $A$ such that, with probability at least $1 - \phi$,

$$\left\| A - C \left( DW \right)^+ D R \right\|_F \leq (1 + \varepsilon) \left\| A - C \left( C^+ A \right) A \right\|_F$$
Relative error CUR

For any $k$, $O(\text{SVD}_k(A))$ time suffices to construct $C$, $U$, and $R$ s.t.

$$\left\| A - C \underbrace{(D W)^+}_{\text{U}} D R \right\|_F \leq (1 + \varepsilon) \| A - A_k \|_F$$

holds with probability at least $1 - \Phi$, by picking

$O(\frac{k \log k}{\gamma^2})$ columns, and

$O(\frac{k \log^2 k}{\gamma^6})$ rows.
## CUR decompositions: a summary

| Authors                          | C: variant of the QR algorithm | R: variant of the QR algorithm | U: minimizes $||A-CUR||_F$ | Notes |
|---------------------------------|--------------------------------|--------------------------------|--------------------------|-------|
| G.W. Stewart                    |                                |                                |                          | No a priori bounds  
Solid experimental performance |
| (Num. Math. '99, TR '04)        |                                |                                |                          |       |
| Goreinov, Tyrtshnikov, &        | C: columns that span max volume | U: $W^+$                       |                          | Existential result  
Error bounds depend on $||W^+||_2 $  
Spectral norm bounds! |
| Zamarashkin (LAA '97, Cont. Math. '01) |                                | R: rows that span max volume |                          |       |
| Williams & Seeger               | C: uniformly at random          | U: $W^+$                       |                          | Experimental evaluation  
A is assumed PSD  
Connections to Nystrom method |
| (NIPS '01)                      |                                | R: uniformly at random          |                          |       |
| D., Kannan, & Mahoney           | C: w.r.t. column lengths        | U: in linear/constant time      |                          | Randomized algorithm  
Provable, a priori, bounds  
Explicit dependency on $A - A_k$ |
| (SODA '03, '04)                 |                                | R: w.r.t. row lengths           |                          |       |
| D., Mahoney, & Muthukrishnan    | C: depends on singular vectors of $A.$  
U: (almost) $W^+$            | R: depends on singular vectors of $C$ |                          | $(1+\gamma)$ approximation to $A - A_k$  
Computable in $SVD_k(A)$ time. |
| ('05, '06)                      |                                |                                |                          |       |
Open problem

Is it possible to construct a CUR decomposition satisfying bounds similar to ours deterministically?


**Main algorithm:** there exist $k$ columns of $A$, forming a matrix $C$, such that the smallest singular value of $C$ is “large”.

We can find such columns in $O(mn^2)$ time deterministically!