

石崎暖人\*, 宮田考史\*

(\*福岡工業大学)

## 1. はじめに

一般にデータ分析で用いられるデータセットは、大規模な縦長行列である。分析の中で重要なことは、大規模なデータを縮小及び圧縮するためにデータを近似することである。

本研究では、行列 $A \in \mathbb{R}^{m \times n}$ (ただし、 $m \geq n$ )を近似するために、CUR分解[1]と呼ばれる行列分解を用いて行列近似する手法を述べる。

## 2. CUR分解

CUR分解は行列 $A$ に含まれている実際の列及び行を用いて行列を分解する方法で、式(1)の形で表される低ランク行列近似である。

$$A \approx CUR. \quad (1)$$

ここで、式(1)中の行列 $C \in \mathbb{R}^{m \times k}$ 及び行列 $R \in \mathbb{R}^{k \times n}$ は、それぞれ行列 $A$ の列の部分行列と行の部分行列である。また、行列 $U \in \mathbb{R}^{k \times k}$ は式(1)の右辺が行列 $A$ に可能な限り近づくように計算され、以下の2つの最小二乗問題を解くことで決定される。ただし、 $k \ll n$ である。

- ① 行列 $X \in \mathbb{R}^{k \times n}$ に対する最小二乗問題 $CX \approx A$ ,
- ② 行列 $U \in \mathbb{R}^{k \times k}$ に対する最小二乗問題 $R^T U^T \approx X^T$ .

CUR分解の最大の利点として、特異値分解と比較すると、CUR分解は $A$ の部分行列を用いるため、例えば、 $A$ が疎行列である場合、特異ベクトル行列は疎性が破壊されるが、 $C$ と $R$ は $A$ と同様に疎になり、 $A$ の構造を維持しやすい。

この $C$ と $R$ を決定するために、 $A$ から列選択及び行選択が必要となる。本研究のCUR分解では、この列及び行選択のために離散経験的補間法(Discrete Empirical Interpolation Method, 以下DEIM)[1]を用いる。

## 3. 離散経験的補間法(DEIM)

DEIMは、次元削減に使われる数値手法で、このDEIMにより最終的に、 $k$ 個の列と行を選択して列番号の集合である $p \in \mathbb{N}^k$ 及び行番号の集合である $q \in \mathbb{N}^k$ を得る。

そのために、まず行列 $A$ の特異値分解 $A = W\Sigma V^T$ を計算し、列選択ベクトル $p$ を得るために右特異ベクトル行列 $V$ に対してDEIMを実行する。また、行選択ベクトル $q$ は左特異ベクトル行列 $W$ に対してDEIMを実行する。

## 4. アルゴリズム

DEIM型CUR分解のアルゴリズムを以下に示す。

Step 1: 特異ベクトルを求めるために行列 $A$ の特異値分解を計算する。

Step 2: 特異ベクトルに対してDEIMを実行する。

ここで選択される列・行の個数は反復回数を $t$ とすると $h = k/t$ 個である。

Step 3: 残差行列 $E$ を計算する。別々に選択する場合の残差行列 $E$ は式(2)及び式(3)のように定義される。

$$E_1 = A - CX, \quad (2)$$

$$E_2 = A - YR. \quad (3)$$

ここで、式(2)は列選択の際で最小二乗問題①を解くことで決定され、式(3)は行選択の際で最小二乗問題 $R^T Y^T \approx A^T$ を解くことで決定される。

また、同時に選択する場合の残差行列 $E$ は最小二乗問題①及び②を解いた上で式(4)のように定義される。

$$E = A - CUR. \quad (4)$$

Step 4: 行列 $E$ に対してStep1から3を $t$ 回繰り返して最

最終的に $k$ 個の列と行を選択する。そして、最小二乗問題①、②により行列 $A$ のCUR分解を決定する。

## 5. 本研究

本研究では、拡大係数行列を用いて特異値問題を固有値問題へ変換後、二分法と逆反復法を組み合わせ一部に必要な特異ベクトルのみを求める。本手法(partial SVD)とすべての特異ベクトルを求める手法(SVD)について近似誤差と計算時間を比較する。

また、列・行選択の際にDEIMを反復させて一度に選択する列及び行の個数を変化させつつ、列と行を異なる反復で別々に選択する方法(Separate selection)と同じ反復で同時に選択する方法(Simultaneous selection)の近似誤差を調べる。

## 6. 数値実験

入力行列 $A$ をSuiteSparse Matrix Collectionから取得した疎行列Maragal\_5を用いて、 $k = 300$ としてCUR分解を行った。図1に近似誤差の比較結果を示し、図2に計算時間の比較結果を示す。ただし、 $A$ の大きさは $4654 \times 3320$ である。ここで、近似誤差を表す指標として2ノルムを用いた相対近似誤差 $\varepsilon$ を式(5)で定義する。

$$\varepsilon = \|A - CUR\|_2 / \|A\|_2. \quad (5)$$

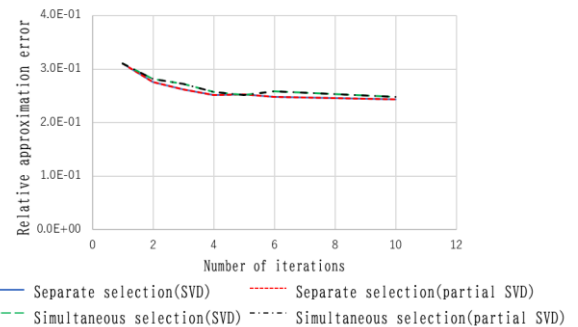


図1 近似誤差

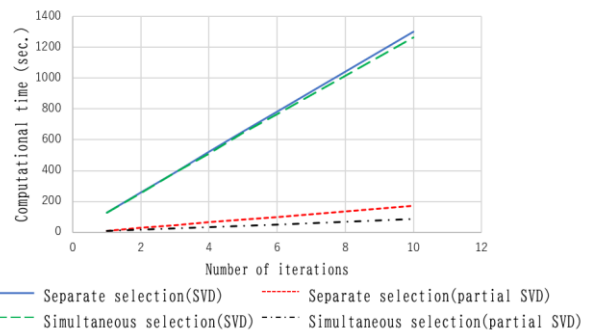


図2 計算時間

図1より、どの手法も同程度の近似誤差を示した。また、図2からSVDより一部のみを計算するpartial SVDを用いることで計算時間を短縮することができた。

## 参考文献

- [1] P. Y. Gidisu, M. E. Hochstenbach, A DEIM-CUR fatocization with iterative SVDs, J. Comput. Math. Data Sci., 12(2024), 100095.