jh250032

QR 分解に関する高性能計算技術の研究

深谷 猛（北海道大学・情報基盤センター）

概要

本研究課題では、基本的な行列計算の一つである QR 分解に対して、様々な観点から研究開発に取り組む。課題参加者は、これまで独立して QR 分解に関連する研究を行った実績があり、本課題では、そのような研究者同士が協力して課題の解決に取り組むことで、QR 分解とその関連技術に関して、新しい技術や知見を創出することを目指す。継続課題の 3 年目となる 2025 年度では、まず、縦長行列の列ピボット付き QR 分解に対して開発したコレスキー QR 型アルゴリズムの実装方法の改良を進め、最新のプログラムの性能評価の結果から、従来手法に対する十分な有効性が確認できた。また、QR 分解に対する乱択アルゴリズムおよびコレスキー QR 型アルゴリズムの GPU 実装を進め、実際に GPU 上での性能評価により、各アルゴリズムの特徴を調査することができた。一方で、当初計画した実施項目のいくつかに関して、十分な進展を得ることができなかった点は、次年度への反省点である。

1 共同研究に関する情報

1.1 共同研究を実施した拠点名

- 北海道大学 情報基盤センター
- 東北大学 サイバーサイエンスセンター
- 東京大学 情報基盤センター
- 京都大学 学術情報メディアセンター
- 大阪大学 D3 センター
- 九州大学 情報基盤研究開発センター

1.2 課題分野

- 大規模計算科学課題分野

1.3 参加研究者の役割分担

- 深谷 猛（北海道大学）：課題代表、全体統括、コレスキー QR アルゴリズム関連の研究開発
- 鈴木 智博（山梨大学）：課題副代表、統括補佐、タイルアルゴリズム関連の研究開発
- 大島 聡史（九州大学）：BLR 行列の QR 分解関連の研究開発、GPU 実装に関する助言
- 伊田 明弘（海洋研究開発機構）：BLR 行列の QR 分解関連の研究開発
- 岩下 武史（京都大学）：QR 分解の応用（線形ソルバー関連）
- 佐竹 祐樹（北海道大学）：QR 分解の応用（行列方程式関連）
- 工藤 侑也（北海道大学・大学院生）：QR 分解の応用（線形ソルバー関連）
- 阿部 龍仁（北海道大学・大学院生）：Randomized アルゴリズムに関する性能評価
- 加藤 勇太（北海道大学・大学院生）：QR 分解の応用（線形ソルバー関連）
- Zhang Peng（山梨大学・大学院生）：タイルアルゴリズム関連
- 久保田 龍（山梨大学・大学院生）：タイル

アルゴリズム関連

2 研究の目的と意義

行列を都合のよい行列の積に分解する計算（行列分解）は、数値線形代数分野の基本技術の一つであり、様々なアプリケーションにおいて利用されるとともに、数値線形代数アルゴリズムを構成する部品としての役割を持つ。そのため、行列分解計算の高性能化（高速化）は重要な課題であり、ハードウェアの特徴を考慮した上で、アルゴリズムから実装方法まで多岐にわたる研究開発が行われている。

本研究課題では、主要な行列分解計算の一つである、行列の QR 分解を扱う。QR 分解は与えられた行列を直交行列と上三角行列の積に分解する計算であり、最小二乗問題が代表的な応用例として知られている。また、行列の固有値計算や特異値計算とも密接な関わりを持っている。加えて、ベクトルの直交化（直交基底の生成）が QR 分解と等価であるため、数値安定性を向上させる目的等で、（ブロック版の）クリロフ部分空間法などにおいても需要がある。したがって、QR 分解の性能を向上させることで、様々な科学技術計算の効率化に貢献できる。

現在、QR 分解に対して異なる特徴を持った多様な数値計算アルゴリズムが存在する。一方、QR 分解の計算が行われる環境も、マルチコア CPU、GPU、分散並列システム（CPU と GPU が混在するヘテロな環境含む）と多種多様である。更に、計算対象の行列も縦長行列から正方行列まで様々な形状があり、加えて、最近では BLR（Block Low Rank）行列の QR 分解のような問題設定も登場している。

このような背景を踏まえた上で、本研究課題では、QR 分解とその関連技術に関して、異なる知識や研究実績を有する研究者を集め、協力して研究開発に取り組む。具体的には、参加する各研究者がこれまでに個別に研究開発をおこなってきた技術や知識を土台として、それらを相互に組み合わせることで、QR 分解とその関連技術の高性能化に資する新しい技術や知見を創出することが大局的な目的である。

本課題は 2023 年度から実施しており、2025 年度は 3 年目である。課題申請の段階では、以下の 6 つの実施項目を計画した。

1. 縦長行列の列ピボット付き QR 分解に対するコレスキー QR 型アルゴリズムの改良
2. 分散並列環境向けタイル QR アルゴリズムの実装と性能評価
3. GPU 向け BLR 行列の QR 分解アルゴリズムの改良
4. QR 分解を必要とする線形計算アルゴリズムへの応用
5. QR 分解に関するライブラリの整備とベンチマークの実行
6. GPU 環境におけるコレスキー QR 型アルゴリズムの実装と性能評価

上記の項目のうち、最初の 5 つは前年度から継続する項目であり、最後の 1 つが 2025 年度に新たに設定した項目である。

3 当拠点公募型研究として実施した意義

JHPCN 課題では、特徴の異なる多種多様な計算機環境を使用可能であり、QR 分解に代表される基本的な線形計算技術の研究開発に適している。例えば、統一的な視点で、各計算機上で、特徴の異なる複数のアルゴリズム（や実装）を評価・分析することで、アルゴリズム開発者とアルゴリズム利用者の双方に有益な知見を得ることができる。また、開発したアルゴリズムをライブラリとして整備・公開することで、各

センターで実行されているアプリケーションの高性能化に貢献できる。更に、JHPCN では、計算科学、データ科学、機械学習などの幅広い応用分野の研究課題が毎年実施されており、それらの課題の関係者とシンポジウム等で交流することができ、基盤技術を扱う本課題にとって大きな意義・利点がある。

4 前年度までに得られた研究成果の概要

1 年目（2023 年度）および 2 年目（2024 年度）に得られた主な成果は以下の通りである。

- 縦長行列の列ピボット付き QR 分解（QRCP）に対して、コレスキー QR 型アルゴリズムを開発した。マルチコア CPU および分散並列環境で性能評価を実施し、既存手法に対して十分な優位性があることを確認した。更に、開発した手法の実装方法の改良について、検討および検証を進めている。
- 非縦長行列の QR 分解に対して、反復型のコレスキー QR アルゴリズムとブロック版の Gram-Schmidt アルゴリズム（BCGS2）を組み合わせた手法を考案し、一次元のデータ分散を採用した分散並列版のプログラムを開発した。北大 Grand Chariot や東大 BDEC（Odyssey）を用いた性能評価を実施し、提案手法の有効性を検証した。
- 最新のマルチコア CPU 環境および分散並列環境において、開発したタイル版の QR 分解プログラムの動作検証を実施した。また、構造が QR 分解よりシンプルなコレスキー分解のアルゴリズムを対象に、分散並列環境でのタイル版アルゴリズムの性能評価等を実施し、QR 分解のタイル並列版プログラムの研究開発に有益となる知見を得た。
- GPU 環境向けの BLR 行列の QR 分解アルゴリズムのプログラムの整理および分析を進めた。また、MIG の利用など GPU の効率的な利用方法など、関連する GPU 利用技術についても、他の JHPCN 課題などと連携して研究を進めた。
- QR 分解に関連する乱択アルゴリズムについて、CPU 環境でプログラムを実装し、初期段階の性能評価を実施し、その有効性や振る舞いを調査した。
- 連立一次方程式に対する低精度演算を用いた混合精度型反復改良法に QR 分解を組み込んだ手法の開発を進めている。また、行列方程式に対する低ランク近似を用いた解法に関して、QR 分解を応用した解法の検討を行っている。
- 本課題に関連して開発している QR 分解のプログラムコードについて、ベンチマークあるいはライブラリとしての公開を念頭に置いた整理・整備を段階的に進めている。

5 今年度の研究成果の詳細

2 節の最後に記載した通り、2025 年度は 6 つの実施項目を計画した。これらの項目のうち、項目 2, 3, 4 に関しては、様々な事情により、当初の計画通りの進展を得ることができなかった。項目 2 に関しては、タイル QR アルゴリズムの最低限の実装は完了したが、性能評価を実施するには至らなかった。項目 3 に関しては、MIG を用いた実装および内部の関数（アルゴリズム）の改良の可能性に関する検討の継続にとどまった。項目 4 に関しては、連立一次方程式に対する混合精度型線形ソルバーの簡単な動作検証を実施した上で、QR 分解を利用した形でのアルゴリズム改良を検討している段階で

ある。また、行列方程式に対する低ランク近似を用いた反復解法のアルゴリズム改良に取り組み、その内部で用いるクリロフ部分空間の正規直交基底の生成に対する QR 分解の応用を検討した。

一方で、項目 1, 5, 6 に関しては性能評価結果を含む一定の成果が得られたので、それらについて以下で詳しく報告する。

5.1 （項目 1）縦長行列の列ピボット付き QR 分解に対するコレスキー QR 型アルゴリズムの改良

縦長行列の列ピボット付き QR 分解に対して、2023 年度に開発したコレスキー QR 型アルゴリズムについて、2024 年度より実装方法の改良を進めており、2025 年度もこれを継続して行った。改良のポイントは、アルゴリズム内部で必要となるグラム行列の対角ピボット付きコレスキー分解の計算手順の変更と、数学的な考察に基づくグラム行列計算の一部の省略である（図 1）。どちらも計算対象の行列の列数（n）が行数（m）に対して相対的に大きくなる場合に効果が大きくなることが期待される。

これらの実装方法の改良を組み込んだ、コレスキー QR 型アルゴリズムの最新のプログラムについて、いくつかの計算機システム上で性能評価を実施した。図 2 は阪大の SQUID 上で実施した性能評価結果の一例である。開発しているコレスキー QR 型アルゴリズムの最新版のプログラムの性能（実行時間）を分散並列計算（強スケーリング）で測定した結果と、代表的な従来手法（ハウスホルダー QR）に対する速度向上を評価した結果を示している。図 2 より、我々のコレスキー QR 型アルゴリズムは並列数（プロセス数）の増加に対して、十分な並列性能を示しており、同時に、ハウスホルダー QR に対して十分な優位性（高速性）が確認できる。

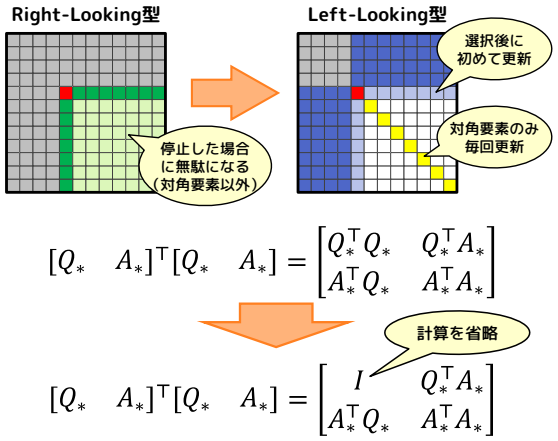

図 1　縦長行列の QRCP に対するコレスキー QR 型アルゴリズムの実装方法の改良：グラム行列に対する対角ピボット付きコレスキー分解の計算手順の変更（上）とグラム行列の計算における演算の省略（下）

今後は、上述の実装方法の改良の効果の詳しい検証や、アーキテクチャの異なる複数の計算機システム間での性能比較などを実施し、開発しているアルゴリズムの特徴や優位性をより明らかにすることを目指す。

5.2 （項目 5）QR 分解に関するライブラリの整備とベンチマークの実行・（項目 6）GPU 環境におけるコレスキー QR 型アルゴリズムの実装と性能評価

QR 分解に対する乱択アルゴリズムの GPU 実装を行い、同時にコレスキー QR 型アルゴリズムについても GPU 実装を行い、最小二乗問題を例題として、両者の性能比較を実施した。

近年、応用数学の分野で乱択アルゴリズムが活発に研究されており、QR 分解（や最小二乗問題）に対する有望なアルゴリズムが提案されている。これを踏まえて、前年度までに CPU 環境において乱択アルゴリズムを実装し、その知見を活かして、2025 年度は GPU 環境向けのプログラム実装を進めた。具体的には、ガウ

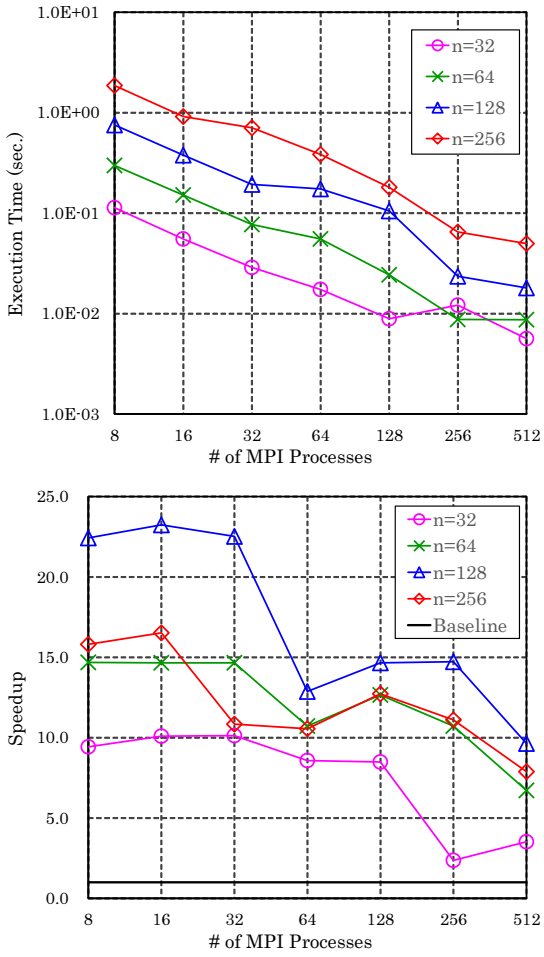

図 2　縦長行列の QRCP に対するコレスキー QR 型アルゴリズムの最新のプログラムに関する性能検証結果の一例：分散並列計算（強スケーリング）における実行時間（上）と代表的な従来手法（ハウスホルダー QR）に対する高速化率（下）、計算機環境：阪大 SQUID、2 プロセス/ノード、38 スレッド/プロセス、$m = 16777216$

シアン行列によるスケッチ行列を用いたアルゴリズムと、SRCT（Subsampled Randomized Cosine Transform）によるスケッチ行列を用いたアルゴリズムを実装した。どちらのアルゴリズムも GPU 環境上での乱数生成器が必要であり、さらに、後者のアルゴリズムでは離散コサイン変換の処理が必要となる。乱数生成器に関しては cuRAND を利用し、離散コサイン変換に関しては、FFT を利用するアルゴリズムを cuFFT のルーチンを使う形で実装した。

同時に、性能比較対象として、コレスキー QR 型アルゴリズム（CholeskyQR2）について、cuBLAS を用いる形で GPU 実装を行った。そして、二種類の乱択アルゴリズムとコレスキー QR 型アルゴリズムと cuSOLVER のルーチン（ハウスホルダー QR に基づく手法と思われる）の 4 種類について、JCAHPC の Miyabi-g（1 ノード）を用いて性能評価を実施した。

図 3 はその結果の一例である。乱択アルゴリズムでは、スケッチサイズと呼ばれるパラメータが存在し、この値に応じて計算結果の精度と計算時間が変化する。そこで、スケッチサイズを変えながら、計算精度と計算時間を測定し、従来法（決定論的手法）と比較を行った。図 3 より、スケッチサイズを大きくすることで、従来法と同程度の計算結果を乱択アルゴリズムにより得られることが確認できる。計算時間に関しては、ガウシアンスケッチを用いる手法はスケッチサイズに対して、計算時間が急激に増加しており、その有効性は限定的である。一方で、SRCT によるスケッチを用いる手法は、スケッチサイズに対する計算時間の増加率が穏やかで、従来法と同程度の精度を得られるスケッチサイズを採用した場合でも、計算時間の点で十分な実用性が期待できることが確認できた。

次の段階として、今回試したスケッチ手法以外の方法を用いたアルゴリズムの実装および、SRCT の実装の改良を検討することが挙げられる。また、テスト行列の条件数などを変えて、各手法の計算精度と計算時間を詳しく評価することも必要である。

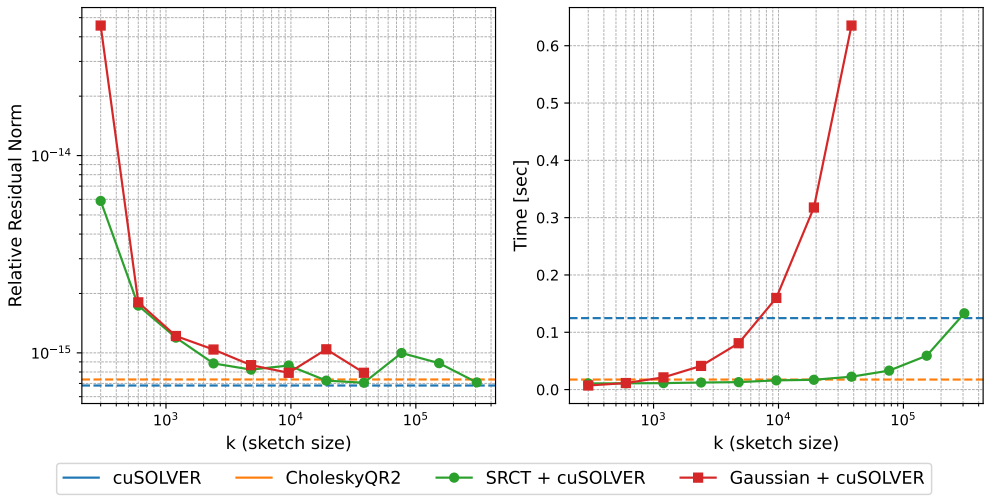

図 3　GPU 環境における最小二乗問題に対する乱択アルゴリズムやコレスキー QR 型アルゴリズムの性能評価結果の一例：計算精度（左）および実行時間（右）、計算機環境：JCAHPC Miyabi-g（1 ノード）、$m = 307201$、$n = 300$

6　進捗状況の自己評価と今後の展望

5 節で記載した通り、2025 年は当初計画で挙げた 6 つの実施項目に関して、一定の成果が得られた項目と成果が限定的であった項目がはっきり分かれた形であった。離散コサイン変換を含む二種類の乱択アルゴリズムの GPU 実装をゼロから行い、同時にコレスキー QR 型アルゴリズムの GPU 実装も実施し、更に GPU 環境で性能評価まで到達できた点は、2025 年度の一番の成果であったと考える。乱択アルゴリズムは理論面の研究が活発であるが、一方で、実装（HPC）面の研究事例が少ない印象であり、日本国内での研究事例の報告も乏しい状況である。それに対して、自分たちでプログラムを開発できたことは、今後、より発展的な性能評価や研究開発を進める上で大きな資産となり得る。

前述の GPU 実装は、主に担当した大学院生の頑張りが大きく、逆に、教員が主担当の項目に関しては、プログラムの開発時間等が十分に確保できず、配分された計算機資源の利用も限定的となってしまった点は大きな反省点である。2026 年度の計画（採択済み）において、この点を反省し、一年でプログラム開発から詳細な性能評価までを計画するのではなく、既にプログラムが概ね開発済み（性能評価の準備が概ね整っている）の項目と、翌年度の性能評価を目指してプログラム開発を進める項目を区別し、前者に対して必要な計算資源を試算する形とした。

また、課題参加者（特に学生以外）が多忙で、本課題の目的の一つである参加者間の議論の機会が 2025 年度は限られていたので、2026 年度は関連する他の JHPCN 課題とうまく連携する形なども視野に入れながら、限られた時間を有効活用して、議論を進めて行けるように努めたい。