2H-08

統計モデルを用いたタイル QR 分解のパラメータチューニング *

霜鳥竜輝 † 鈴木智博 ‡
山梨大学大学院 §

1 はじめに

密行列の数値線形代数において，行列分解はさまざまな計算の前/後処理に使われる重要なアルゴリズムである．近年の科学技術計算の大規模化，高速化の要請に応えるために，高並列向きの行列分解アルゴリズムが求められている．行列分解に対するタイルアルゴリズムは行列を小行列 (タイル) に分割し，個々のタイルに対して処理を行うことで細粒度のタスクを大量に生成できるため，高並列環境向きの行列分解アルゴリズムとして注目されている．このタイルアルゴリズムは，特定の内部パラメータによってその計算性能が大きく変わるため，これらのパラメータを適切に選択する必要がある．今回は，パラメータチューニングのために計算速度の統計モデルを作成した．その結果，モデルの説明変数の自動選択及びチューニング時間の短縮を実現したため，それを報告する．

2 タイル QR 分解

QR 分解は数値線形代数計算において重要な役割を果たす基本的なアルゴリズムである．$m \times n (m \geq n)$ 行列 A の QR 分解は，以下のように定義される．

$$A = QR \tag{1}$$

ここで，Q は $m \times m$ の直交行列，R は $m \times n$ の上三角行列である．数値計算ライブラリ LAPACK[1] の QR 分解ではブロックアルゴリズムが採用されている．ブロックアルゴリズムは，行列を縦方向に分割し，分解と後続行列更新を縦ブロック毎に繰り返すことで行列全体を分解する．しかし，ブロックアルゴリズムは fork-join 型並列実行モデルに基づくため，高並列計算資源を効率よく使用することができない．

行列分解に対するタイルアルゴリズムは，対象とする行列を小行列に分割し，基本演算 (カーネル) を 1 または 2 タイル毎に行う (図 1)．タイルアルゴリズムは細粒度のタスクを大量に生成ことが可能であり，これらを非同期に実行することで高並列計算資源を有効に活用できるアルゴリズムとして注目されている．

3 パラメータチューニング

タイル QR 分解には調整可能な内部パラメータが存在し，このパラメータが適切に選択された場合とそうでない場合では性能が大きく異なる．タイル QR 分解における調整可能なパラメータは，タイルサイズ nb と，内部ブロックサイズ ib である．内部ブロックサイズとは，カーネルにブロックアルゴリズムが用いられるため，その時の縦ブロックサイズのことを指す．

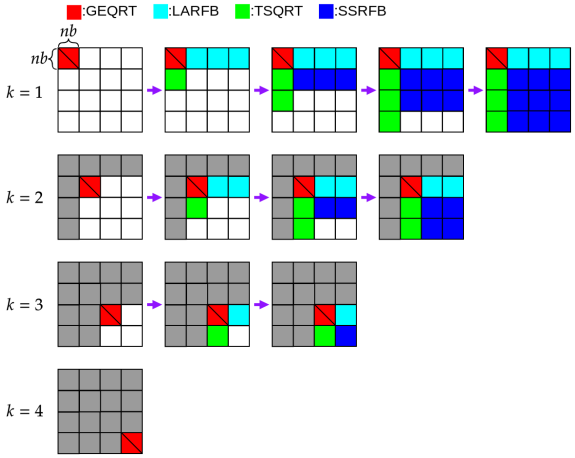


図 1: タイル QR 分解

タイル QR 分解は，SSRFB と呼ばれるカーネルが全体の大部分を占める．SSRFB カーネルは L3BLAS ルーチンが主要の演算であるため，タイルサイズ nb が大きいほど高い性能を発揮する．しかし，nb を大きくしすぎるとタスクの数が少なくなり，並列計算資源に負荷不均衡が生ずる．使用するコア数などの計算資源に対して十分なタスク数を供給しつつ，可能な限り大きなタイルサイズを選択する必要がある．

前述の通り，タイル QR 分解は SSRFB カーネルが処理の大部分を占めていることから，全体性能の向上は SSRFB カーネルの速度向上によってなされる．そのため，SSRFB カーネルの計算速度を最大化するような内部ブロックサイズ ib が求められる．

今回は，パラメータチューニングを行うために，タイル QR 分解，SSRFB カーネルの速度の統計モデルを作成した．統計モデルの生成やそのモデルを用いた最適パラメータの探索について次で述べる．

4 統計モデル

4.1 統計モデルの定義

パラメータチューニングにあたって，今回は 2 種類のモデル $f(n, nb)$ 及び $g(nb, ib)$ を生成した．両モデルの詳細は以下の通りである．なお，これ以降では行列 A を $n \times n$ の正方行列として扱う．

- $f(n, nb)$
  説明変数 $n, nb, n^2, n \times nb, nb^2, n^3, n^2 \times nb, n \times$

---

* Parameter tuning for tile QR decomposition using statistical model
† Ryuki Shimotori
‡ Tomohiro Suzuki
§ University of Yamanashi

nb^2, nb^3

目的変数 タイル QR 分解 1 回あたりの計算速度 $[Gflops]$

サンプル

- $10000 \leq n \leq 60000$(10000 刻み)
- $128 \leq nb \leq 1280$(128 刻み)
- $g(nb, ib)$ に最適化アルゴリズムを適用して取得した最適な ib

- $g(nb, ib)$

説明変数 $nb, ib, nb^2, nb \times ib, ib^2, nb^3, nb^2 \times ib, nb \times ib^2, ib^3$

目的変数 $SSRFB$ カーネル 1 回あたりの計算速度 $[Gflops]$

サンプル

- $n = 5 \times nb$
- $128 \leq nb \leq 1280$(128 刻み)
- $ib = nb/2^x (1 \leq x \leq 6)$

行列分解では計算量がパラメータの 3 次のオーダーに，データ移動がパラメータの 2 次のオーダーに比例する．そのため，n と nb(nb と ib) で構成される分数を含まない 1 次から 3 次の項をすべて列挙したものを説明変数として選定した．

最適な nb は行列サイズに，最適な ib はタイルサイズにそれぞれ依存するため上記のようにモデルを 2 種類に分けた．ポイントは，モデルを 2 種類に分けることによって SSRFB カーネルのデータだけで ib のパラメータ探索ができる点である．SSRFB カーネルの計算速度はタイル QR 分解よりも短時間で計測できるため，モデルを生成するためのデータ取得時間を大幅に短縮できる．すなわち，チューニング時間そのものの短縮につながる．

4.2 統計モデルの生成手法

今回は，Lasso 回帰を用いて統計モデルを生成した．Lasso 回帰の定義は以下．

$$\min_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_i^{(j)} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad (2)$$

ここで，y_i を目的変数，$x_i^{(1)}, \cdots, x_i^{(p)}$ を説明変数，推定したいパラメータの数を p，回帰係数を β_j，N をサンプルサイズとする．Lasso 回帰は最小二乗関数に正則化項 $\lambda \sum_{j=1}^{p} |\beta_j|$ を加えることで，不要と判断された説明変数の係数を 0 にする手法である．これは，モデルの選択と同時に説明変数の数を削減し，説明変数の選択を自動で行う手法であるとも言える．なお，正則化項に含まれる値 λ は正則化の強さを決めるパラメータであり，この値を調節して説明変数の数を増やしたり減らしたりする．今回は $f(n, nb)$，$g(nb, ib)$ についてそれぞれ説明変数が 1 個から 9 個の 9 種類ずつのモデルを生成し，その中から最適なモデルを選択した (「最適なモデル」の基準は後述)．Lasso 回帰や λ の調節は，R 言語の glmnet パッケージ [2] を用いることで自動で実行できる．

5 最適パラメータ

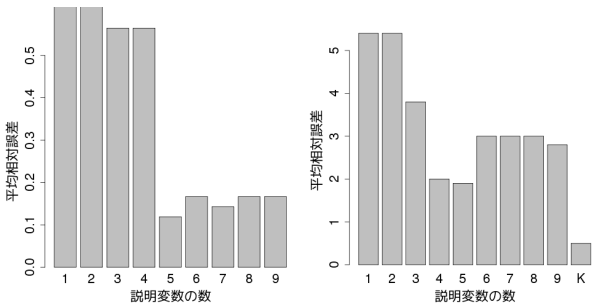
与えられた行列サイズ n に対して，モデル $f(n, nb)$ を使って最適な nb を取得し，この nb に対して $g(nb, ib)$ を使って最適な ib を取得する．最適パラメータを取得するために，今回は R 言語の optimize 関数を用いた．

$f(n, nb)$ と $g(nb, ib)$ についてそれぞれ 9 種類ずつのモデルを作っているため，最適解も 9 個ずつ求められるはずである．これらの 9 個ずつの最適解のうち，実際の最適パラメータに最も近いものを最適なパラメータとし，それを導出したモデルを最適なモデルとする．

モデルの予測性能を評価するために，モデルによって予測された最適パラメータと実際の最適パラメータの平均相対誤差を取得する．そして，その平均相対誤差が最も小さいモデルを「最適なモデル」とした．

横軸に説明変数の数，縦軸に平均相対誤差をプロットした棒グラフを図 2 に示す．なお，SSRFB カーネルの性能モデルは Kurzak[3] によって既に考案されている (式 (3))．比較のために，図 2b には式 (3) で得られた最適パラメータに関する相対誤差も合わせて記載する (軸ラベルを「K」と表記する)．

$$\text{計算速度} \sim \frac{\beta_1}{\beta_2 + \beta_3 \frac{ib}{nb} + \beta_4 \frac{1}{ib} + \beta_5 \frac{1}{nb}} \qquad (3)$$



(a) $f(n, nb)$ **(b)** $g(nb, ib)$

図 2: 最適解の相対誤差

図 2 から，$f(n, nb)$，$g(nb, ib)$ のいずれも説明変数が 5 つのモデルが最適なモデルであると言える．しかし，Kurzak のモデルと比較すると平均相対誤差が大きい．本来，説明変数に用いたパラメータは計算時間を表しているが，$f(n, nb)$，$g(nb, ib)$ はこれらのパラメータで計算速度を表そうとしているため，真の最適パラメータが導出できなかったと考えている．

6 おわりに

今回は統計モデルを用いてパラメータチューニングを行った．この時，モデルを $f(n, nb)$，$g(nb, ib)$ と分けることでチューニング時間を短縮できること，Lasso 回帰を用いることで説明変数の自動選択が行えることを確認した．しかし，予測性能は既に考案されているモデルに及ばなかった．

今後の課題として，最適パラメータの予測性能の向上があげられる．今回説明変数として用いたパラメータは計算時間を表すため，これらのパラメータを計算量 (タイル QR 分解なら n^3，SSRFB カーネルならば nb^3) で割ることで計算速度を表現し，この計算速度でモデル生成を行えばより予測性能の高いモデルが取得できると考えている．

参考文献

[1] http://www.netlib.org/lapack/
[2] https://cran.r-project.org/web/packages/glmnet/glmnet.pdf
[3] J. Kurzak, "Modeling and Tuning Parallel Performance in Dense Linear Algebra" (2008)