レゾルベントの作用の計算に混合精度による残差反復法を用いたフィルタ対角化法の実験

村上 弘[1,a)]

概要：実対称定値一般固有値問題の固有対で固有値が指定された区間にあるものをフィルタ対角化法を用いて求めることにする．問題の係数行列は倍精度で与えられているとする．フィルタとして単一のレゾルベントから構成される簡易な多項式型のものを用いることにする．そうして単一のレゾルベントの作用に対応するシフト行列の連立 1 次方程式は，単精度の数値として求めたシフト行列の分解結果を用いて残差反復法で解くことにする．これにより行列分解の結果を保持するための記憶量を半分に減らすことができる．

キーワード：フィルタ対角化，固有値問題，レゾルベント，固有対，反復改良，正規直交化

Experiments of Filter Diagonalization Method Which Uses Mixed Precision Iterative Refinement Method to Calculate the Action of a Resolvent

Hiroshi Murakami[1,a)]

Abstract: By the use of the filter diagonalization method, we solve those eigenpairs of a real symmetric definite generalized eigenproblem whose eigenvalues are in a specified interval. And we assume both coefficient matrices of the problem are given in double-precision. We are to use a simple polynomial type filter which consists of a single resolvent. We solve the simultaneous linear equations, that corresponds to the action of the resolvent, by the mixed precision iterative refinement method which uses matrix factors calculated and stored in single-precision. By that we can reduce the amount of the storage to hold the matrix factors to half.”

Keywords: filter diagonalization, eigenproblem, resolvent, eigenpair, iterative refinement, orthonormalization

1. はじめに

以前の報告 [11][12] に対して今回の報告で追加する内容は，フィルタに用いるレゾルベントの作用を与える大規模な連立 1 次方程式の解法に対して残差反復法を導入してそれを混合精度を用いて計算したことである．行列分解や前進後退代入の計算を「通常精度」よりも「低精度」の数値と演算を用いて行うが，残差を改良する反復を 2~ 3 回程度行なうことにより連立 1 次方程式の解を「通常精度」で求めることができる．このような変更を計算方法に加えて実験を行い，その例を示している．

フィルタ対角化法を用いて，与えられた実対称定値一般固有値問題 $A\mathbf{v} = \lambda B\mathbf{v}$ の固有対 $(\lambda, \mathbf{v})$ で指定された区間 $[a, b]$ に固有値 λ があるものを一斉に近似して求める．

今回の実験に用いたフィルタ $\mathcal{F}$ は，単一のレゾルベント $\mathcal{R}(\rho) \equiv (A - \rho B)^{-1}B$ で構成されたものであり，式 (1) のシフトが実数 ρ のレゾルベントの実多項式 P の形であるか，あるいは式 (2) のシフトが虚数 ρ' のレゾルベントの虚部の実多項式 P の形であるかのいずれかであるとする [10].

$$\mathcal{F} \equiv P(\,\mathcal{R}(\rho)\,)\,. \tag{1}$$

$$\mathcal{F} \equiv P(\,\mathrm{Im}\,\mathcal{R}(\rho')\,)\,. \tag{2}$$

レゾルベント $\mathcal{R}(\rho)$ のベクトル $\mathbf{x}$ への作用 $\mathbf{y} \leftarrow \mathcal{R}(\rho)\,\mathbf{x}$

1 東京都立大学数理科学専攻
Department of Mathematical Sciences, Tokyo Metropolitan University, Hachioji, Tokyo 192–0397, Japan
a) mrkmhrsh@tmu.ac.jp

は，シフト行列 $C(\rho) \equiv A - \rho B$ を係数とする連立 1 次方程式 $C(\rho)\,\mathbf{y} = B\,\mathbf{x}$ を解いて実現されるものであり，この連立 1 次方程式の求解の計算がフィルタを適用する処理の主要部である．本報告ではこのレゾルベントの作用を与える大規模な連立 1 次方程式は直接法で解くことを前提にする．その場合は行列分解に掛かる演算量と特に分解結果を保持するための記憶量が計算実施上の制約になりがちである．

大規模問題の場合に，単一のレゾルベントから構成されるフィルタは複数から構成されるものと比べて，行列分解の演算量と分解結果を保持するために必要な記憶量が少ないことが大きな利点であるが，その反面としてレゾルベントのシフトや多項式を調整しても遮断特性の急峻さや通過域での伝達率の均一性をあまり良くできないことがある．

一般にフィルタの特性が良くないと，得られる固有対の近似も良くならない．そこで「特性の良くないフィルタ」の適用を繰り返して固有対の近似を改良しようとする．同一のフィルタを繰り返し用いるときは，レゾルベントの作用を与える連立 1 次方程式の係数行列は変わらないので，最初に 1 度だけ構成した行列分解を使い続けることができる．しかし通過域に於いて伝達率の均一性が良くないフィルタの適用を単純に反復すると，反復に伴って「求めたい固有ベクトル」の伝達率の相互の比は 2 乗，3 乗となり拡大する．扱える数値の有効桁数に限界のある計算では，反復した結果のベクトルに含まれる「求めたい固有ベクトル」の情報の精度は，その伝達率が相対的に小さいものほど低下あるいは全く失われる．この反復に伴う「求めたい固有ベクトル」の情報の精度低下を抑制するために，フィルタだけを単純に反復するのではなくて，各回のフィルタの適用のたびにベクトルの組に対して B-正規直交化を施すことにする [11], [12], [13]（付録 A.4 参照）．これは構造解析の分野では「(直交化付き）同時逆反復法」と呼ばれる方法（文献 [4], [5], [6], [7]）の類似である（その原理は「直交化（同時）反復法」（Orthogonal Iteration）[9] である）．

本報告の実験では，一様乱数を元にして作られた B-正規直交化ベクトルの組から始めて，「フィルタを適用した後に B-正規直交化を施す操作」を数回反復することで近似不変部分空間の基底を改良する．そうしてその反復により得られた基底から改良された近似固有対をうまく抽出する．ここで B-正規直交化とは，与えられたベクトルの組から正定値行列 B をベクトルの内積の重みとする B-正規直交基底を構成する方法のことである．今回の実験には B-正規直交化として（閾値による切断を入れた）特異値分解（B-SVD）を用いた．

1.1 B-正規直交化法について

m 個の列ベクトルを並べた組 Y の B-正規直交化には，（閾値による切断を入れた）B-特異値分解（B-SVD）を利用した．その計算方法を示す．

1) m 次の対称正定値行列 $G \Leftarrow Y^T B Y$ を作る．

2) Rutishauser の Jacobi 法 [2] を用いて固有値分解 $G \Rightarrow UDU^T$ の m 次の対角行列 D と直交行列 U を求める（D の対角は減少順にとる）．

3) $Y \Leftarrow YU$ は B-直交する列ベクトルの組．

元の Y が悪条件であったり，G を作る際の数値丸め誤差の蓄積などにより，上記の計算で得られる Y の列は B-直交性が不十分になることがある．そこで B-直交性の改良を狙って，上記 1) から 3) までを高々 2，3 回繰り返し，途中で G がほぼ対角になれば，B-直交性が十分になったとして繰り返しから抜ける．そうして Y の列で対応する G の対角要素の平方根（ノルム）が閾値未満のものは棄てる．棄てられずに残った各列を集めてそれぞれに対応するノルムの逆数を乗じて新たに作った Y は B-正規直交化ベクトルの組になる．このようにして得られた Y の列の数は，元の Y の列の数から閾値による切断で棄てられた列の数だけ少なくなる．

注意として，ベクトルの組にフィルタを適用する際には，その組を任意に分割してそれぞれ独立並行に処理をすることが可能であるが，間に B-直交化の処理を挟むとその段階は必要なベクトルが揃うまで完了できずに待つことになり，そこで処理に同期が発生するので，並列分散処理用には不便になることである．また，ベクトルの組に対する直交化の処理は線形の作用ではないので，直交化の操作を途中に挟むと処理全体としての作用も線形ではなくなり，処理全体の伝達特性を数式で表せなくなる．

2. フィルタ対角化法の概略

行列 A，B が実対称で B が正定値の一般固有値問題 (3) の固有対 $(\lambda, \mathbf{v})$ であって，固有値 λ が指定された区間 $[a, b]$ にあるものだけをうまく近似して求める．

$$A\mathbf{v} = \lambda B\,\mathbf{v} \tag{3}$$

フィルタはそのためにうまく調整された線形作用素であり，固有値が区間内にある固有ベクトルは良く通過させるが，固有値が区間から離れた固有ベクトルはなるべく阻止するように構成する．するとフィルタは「必要な固有ベクトルで張られた不変部分空間」への射影作用素の近似になる．そこで線形独立性の良いベクトルを十分多く生成してフィルタを作用させることで像ベクトルを作り，それらの線形結合をうまく選ぶことで不変部分空間の近似空間の基底を構成する．そのようにして得られた基底に Rayleigh-Ritz 法を適用することで必要な固有対の近似が得られる．

3. 実験に用いたフィルタ

今回の実験で用いるフィルタは単一のレゾルベントの多項式型とし，その多項式としてチェビシェフ多項式を用いた．このような簡易な構成のフィルタは伝達特性をあまり良くすることができない．

求めたい固有対の固有値の区間 $[a, b]$ が固有値分布の下端である場合は，シフト ρ を実数に選ぶことができて，フィルタは以下の式 (4) により与えられる．

$$\mathcal{F} \equiv g_{\mathrm{s}}\, T_n(\, 2\gamma\, \mathcal{R}(\rho) - I\,)\,. \tag{4}$$

あるいは固有値の区間 $[a, b]$ の位置を自由に設定したい場合にはシフト ρ' を虚数に選ぶことで，フィルタは以下の式 (5) により与えられる．

$$\mathcal{F} \equiv g_s\, T_n(\,2\gamma' \,\mathrm{Im}\, \mathcal{R}(\rho') - I\,)\,. \qquad (5)$$

ここで $T_n(x)$ は x の n 次のチェビシェフ多項式，I は恒等作用素，そうして Im は虚部をとり出す作用素を表す．また γ や γ' は実数の定数であり，g_s は阻止域に於ける伝達関数の大きさの上限である（より詳しい構成法は付録の副節 A.1 に記述した）．

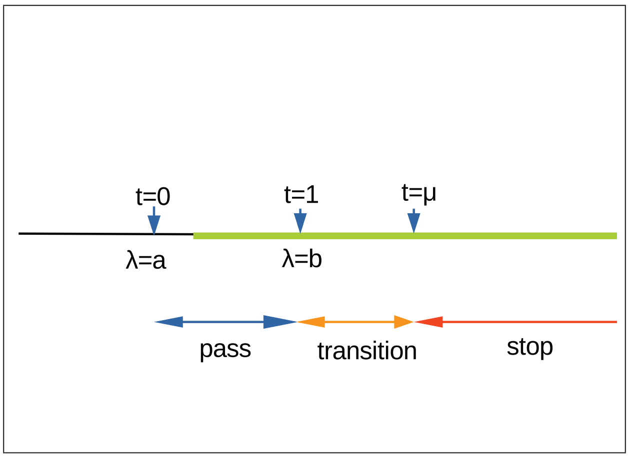

図 1　固有値 λ と正規化座標 t の関係（下端の固有値用の場合）通過域 $t \in [0, 1]$；遷移域 $t \in (1, \mu)$；阻止域 $t \in [\mu, \infty)$．緑色の太線部分は固有値が存在しうる領域

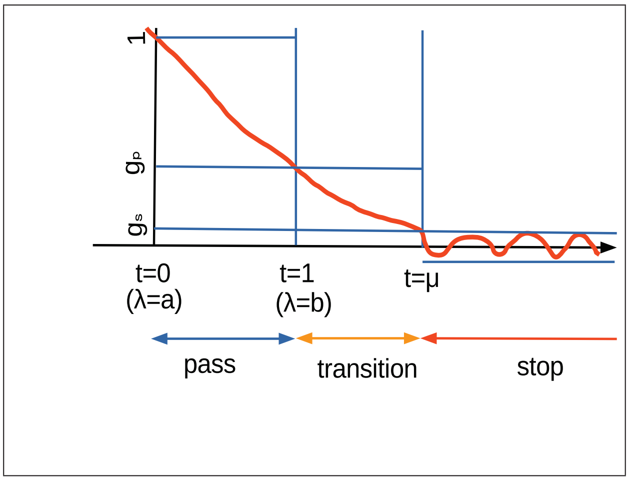

図 2　伝達関数 $g(t)$ の概形（下端の固有値用の場合）

4. 混合精度によるレゾルベントの作用の計算

与えられたベクトル $\mathbf{x}$ に対して，シフトが ρ のレゾルベント $\mathcal{R}(\rho)$ を作用させた結果のベクトル $\mathbf{y}$ を求める計算 $\mathbf{y} := \mathcal{R}(\rho)\mathbf{x}$ は，一般固有値問題の係数の実対称行列を A と B とするとき，$\mathbf{x}$ を与えて対称行列 $C = A - \rho B$ を係数とする連立 1 次方程式 $C\mathbf{y} = B\mathbf{x}$ を $\mathbf{y}$ について解くことである．係数行列 C はシフト ρ が実のときは実対称，ρ が虚数のときは複素対称になる．

本報告の中では A，B，C，ρ，$\mathbf{x}$，$\mathbf{y}$ に通常用いている数値や演算の精度のことを「通常精度」と呼ぶことにする．そうして「通常精度」よりも精度の低い数値や演算のことを「低精度」と呼ぶことにする．

連立 1 次方程式 $C\mathbf{y} = B\mathbf{x}$ を「通常精度」だけを用いて解く場合は，対称行列 C の改訂コレスキ分解 $C =: LDL^T$ を既に構成していれば，まず右辺 $\mathbf{u} := B\mathbf{x}$ を計算して，それに対して C の分解結果を用いた前進後退代入を行うと $\mathbf{y}$ が求まる．つまり実際には C の逆行列を作らずに分解を利用して前進後退代入の操作で C^{-1} の作用を実現しているが，それを便宜上 $\mathbf{y} := C^{-1}\mathbf{u}$ と表すことにする．

つまり，まず $C := A - \rho B$ を作り，C の改訂コレスキ分解を構成して保持しておく．そうしてベクトル $\mathbf{x}$ に対してレゾルベント $\mathcal{R}(\rho)$ を作用させるときは，まず $\mathbf{u} := B\mathbf{x}$ を作り，それから前進後退代入により $\mathbf{y} := C^{-1}\mathbf{u}$ を作る．これがレゾルベントの作用を「標準精度」を用いて適用する通常の計算方法である．

同じ計算を「低精度」の計算で残差反復法を用いて求めるには以下のようにする．いま $A - \rho B$ の「標準精度」の各成分を「低精度」に丸めたものを成分とする対称行列を C_L とする．そうして，あらかじめ C_L を「低精度」の計算で改訂コレスキ分解 $C_L =: L_L\, D_L\, L_L^T$ を構成して保持する．そうして，この分解結果を用いて「低精度」の計算で前進後退代入を行うことにより，C_L の逆を「低精度」のベクトル $\mathbf{r}_L$ に対して作用させる処理のことを，前と同様にここでは $(C_L)^{-1}\mathbf{r}_L$ と表すことにする．すると，レゾルベント $\mathcal{R}(\rho)$ を「通常精度」のベクトル $\mathbf{x}$ に適用してその結果である「通常精度」のベクトル $\mathbf{y}$ を得る計算は，連立 1 次方程式を残差反復法（iterative refinement）を用いて解くことにより，**図 3** のように書ける [1], [3]．ここで B，C，$\mathbf{x}$，$\mathbf{u}$，$\mathbf{y}$ は「標準精度」の行列やベクトル，C_L は C を低精度に丸めた行列，$\mathbf{r}_L$ は低精度のベクトルをそれぞれ表している．ここで ℓ は残差反復法の中の反復回数であり，$\ell = 1$ の場合は単に C^{-1} を「低精度」の C_L^{-1} で置き換えたのと同じ結果になる．

1	:	$\mathbf{u} := B\,\mathbf{x}$;
2	:	$\mathbf{y} := \mathbf{0}$;
3	:	$\mathbf{r}_L := \mathbf{u}$;
4	:	LOOP for k from 1 to ℓ do
5	:	$\mathbf{r}_L := (C_L)^{-1}\,\mathbf{r}_L$;
6	:	$\mathbf{y} := \mathbf{y} + \mathbf{r}_L$;
7	:	if $(k = \ell)$ exit LOOP ;
8	:	$\mathbf{r}_L := \mathbf{u} - C\,\mathbf{y}$;
9	:	end LOOP

図 3　残差反復法によるレゾルベントの作用 $\mathbf{y} := \mathcal{R}(\rho)\,\mathbf{x}$ の計算

図 3 の算法において：
- ステップ 1 は，「通常精度」での演算と代入である．
- ステップ 2 は，「通常精度」での零の代入である．

- ステップ 3 は，右辺の「通常精度」の値を代入に際して低精度に丸めている．
- ステップ 5 は，すべて「低精度」の数値と演算で行なわれ，連立 1 次方程式を前進後退代入により in-place で計算しているところである．
- ステップ 6 は，右辺の中で「低精度」の数値を「通常精度」に変換して加算を行い，その結果を代入している．
- ステップ 7 は次のステップ 8 で残差を求めている処理を一番最後の繰り返しでは省略するためのものである．もしも最終結果の残差が欲しい場合には，この脱出のためのステップを削除する．
- ステップ 8 では，方程式の残差である右辺を「通常精度」の演算で求めてその結果を「低精度」に丸めて代入している．ここで残差の大きさが指定された閾値以下かを調べて繰り返しを中断するようにもできる．

この残差反復法により計算を行うと，解くべき連立 1 次方程式の係数行列の分解は「低精度」のものだけでよくなり，また特に分解後の行列の格納に必要な（語数ではなくて byte 単位で測った）記憶量を少なくできることが利点である．またもしも，「低精度」の方が「標準精度」に比べて使用する計算システム上での演算処理がかなり高速であれば，この残差反復法を用いて精度を混合する方法は計算速度の面においても有利になる可能性がある．

固有値分布から離れているようにシフト ρ を選んでいることから，「標準精度」のシフト行列 $C=A-\rho B$ の逆行列は「低精度」の C_L の逆行列により良く近似されるので，いまの場合の残差反復法はとてもうまく機能する．

なお上記の図 3 の式中では列ベクトル $\mathbf{x}$ を 1 つ与えて列ベクトル $\mathbf{y}$ を 1 つ求めるという形で書いているが，行列 C や B を変えることなく列ベクトルを複数与えてそれらに対応する複数の列ベクトルを一斉に求めるのには，式中で列ベクトルが現れている箇所すべてを列ベクトルを複数並べた行列の形にそれぞれ置き換えるだけでよい．

5. 計算実験について

5.1 一般固有値問題の例題

計算実験に用いた例題の実対称定値一般固有値問題 $A\mathbf{v}=\lambda B\mathbf{v}$ は，1 辺の長さ π の 3 次元立方体を領域として，その表面に於いて零ディリクレ境界条件を課した（符号を逆にした）ラプラス作用素の固有値問題 $-\Delta\Psi=\lambda\Psi$ を有限要素法で離散化して得られるものである．有限要素法の各要素は，元の立方体領域を各辺方向にそれぞれ N_1+1, N_2+1, N_3+1 に等分して生じる直方体である．そうして各要素内での関数展開の基底には最も低次である 3 重線形関数を採用した．この離散化から導かれる一般固有値問題の 2 つの係数行列 A と B の次数はそれぞれ $N=N_1N_2N_3$ となる．また基底に適切な番号付けを行うことで係数行列は下帯幅が $w_L=1+N_1+N_1N_2$ の帯行列にできる（ここで要素分割の数は昇順 $N_1\le N_2\le N_3$ であるとしておく）．各行列の帯内部は実際には極めて疎であるが，計算上は帯内が密であるように扱っている．この例題の固有値はすべて正であり，厳密値を簡単な数式の計算で求めることもできる（付録の副々節 A.2.1 を参照）．

計算で求めた近似固有対 $(\lambda,\mathbf{v})$ の品質評価には，2-ノルムによる相対残差の値を用いた（付録の副節 A.3 参照）．

5.2 実験に用いた 4 通りのフィルタの設定

各実験に用いた 4 通りのフィルタ（R1，R2，C1，C2）は以下のように設定した．

- 固有値が固有値分布の下端付近の区間 $[a,b]=[0,100]$ にある固有対を求めるために用いたフィルタは以下の 2 通りで，シフトが実数の単一のレゾルベントから構成されるものである．
  フィルタ R1 の構成は $n=10$, $\mu=1.5$, $g_{\mathrm{s}}=$ `1E-10` とする．通過域での最小伝達率は $g_{\mathrm{p}}=$ `1.69E-6` になる．
  フィルタ R2 の構成は $n=15$, $\mu=1.5$, $g_{\mathrm{s}}=$ `1E-12` とする．通過域での最小伝達率は $g_{\mathrm{p}}=$ `4.17E-7` になる．
  どちらの場合にも，最初に与えるランダムなベクトルの数は $m=800$ とした．
- 固有値が固有値分布の中間の区間 $[a,b]=[100,200]$ にある固有対を求めるために用いたフィルタは以下の 2 通りで，シフトが虚数の単一のレゾルベントから構成されるものである．
  フィルタ C1 の構成は $n=10$, $\mu=1.5$, $g_{\mathrm{s}}=$ `1E-10` とする．通過域での最小伝達率は $g_{\mathrm{p}}=$ `9.33E-5` になる．
  フィルタ C2 の構成は $n=12$, $\mu=1.5$, $g_{\mathrm{s}}=$ `1E-14` とする．通過域での最小伝達率は $g_{\mathrm{p}}=$ `7.52E-7` になる．
  どちらの場合にも，最初に与えるランダムなベクトルの数は $m=1,300$ とした．

レゾルベントの作用を与える連立 1 次方程式を残差反復法により「低精度」計算を用いて解く場合には，その中の反復回数を ℓ で表す．近似対の改良のためにフィルタを反復する回数は `IT` で表す．

5.3 実験の計算環境

実験に用いた計算機システムは東京大学情報基盤センターの Oakforest PACS のノード（Fujitsu PRIMERGY CX1640 M1）1 つである．ノード内の CPU は 1 つで，CPU は Intel Xeon Phi 7250（クロック周波数は 1.4GHz，コア数 68，32MB L2 キャッシュ）で，理論ピーク性能は 3.05TFLOPS（倍精度）である．主記憶は 6 チャンネルの DDR4-2400 メモリで合計容量が 96Gbyte，それと MC-DRAM が 16Gbyte である．バッチ処理の JOB-CLASS には `regular-cache` を用いたが，これは MCDRAM を主記憶の DDR4 メモリのキャッシュとして動作させるモードになる．プログラムはすべて Fortran90 言語に OpenMP のディレクティブを加えて記述した．コンパイラは Intel Fortan version 19.0.5.281 で，オプションとして `"-Ofast -xMIC-AVX512 -align array64byte -qopenmp"` を指定した．実行時の使用スレッド数として CPU のコア数の 3 倍の値 204 を指定した．用いている線形計算の手法は level-2 BLAS 的であるので，経過時間は大幅に改善できる余地があるように思われる．

6. 解を倍精度で求めた実験の例

IEEE754 の倍精度（2 進 64 ビット，有効 15.95 桁）で各フィルタ（R1，R2，C1，C2）を IT 回反復して（付録 A.4 参照），一般固有値問題の固有対を求めた計算例を示す．

例題とした有限要素法の要素分割は $(N_1, N_2, N_3) = (50, 60, 70)$ であり，それによる一般固有値問題の係数行列の次数は $N = 210,000$，下帯幅は $w_L = 3,051$ である．

固有値 λ が固有値分布の下端の区間 $[0, 100]$ にある固有対はフィルタ R1 と R2 を用いて求め，中間の区間 $[100, 200]$ にある固有対はフィルタ C1 と C2 を用いて求めた．固有値が下端の区間 $[0, 100]$ にある固有対の数は 402 であり，中間の区間 $[100, 200]$ にある固有対の数は 801 である．

最初に与えるランダムなベクトルの数は，固有値が下端の区間にある固有対を求める場合は $m = 800$ とし，中間の区間にある固有対を求める場合は $m = 1,300$ とした．

6.1 倍精度だけを用いた計算の例

比較のためにまず，倍精度だけを用いた計算の結果を示す．フィルタ（R1，R2，C1，C2）を用いて得られた近似固有対のうち固有値が区間 $[a, b]$ にあるものの相対残差の最大値を**表 1** に示し，またそれらをグラフにプロットしたものを**図 4** に示す．さらに経過時間を**表 2** に示す．

表 1 最大の相対残差，各フィルタ（倍精度だけ）

IT	フィルタ R1	フィルタ R2	フィルタ C1	フィルタ C2
1	1.5E-02	1.1E-03	1.5E-03	8.7E-03
2	1.1E-06	2.3E-09	4.2E-10	9.4E-14
3	1.6E-10	3.4E-11	4.2E-14	1.2E-13
4	9.1E-13	1.1E-12	4.1E-14	4.1E-14

表 2 経過時間 (秒)，各フィルタ（倍精度だけ）

IT	フィルタ R1	フィルタ R2	フィルタ C1	フィルタ C2
1	423	535	1,131	1,292
2	695	910	1,964	2,135
3	905	1,228	2,450	2,823
4	1,104	1,544	3,155	3,513

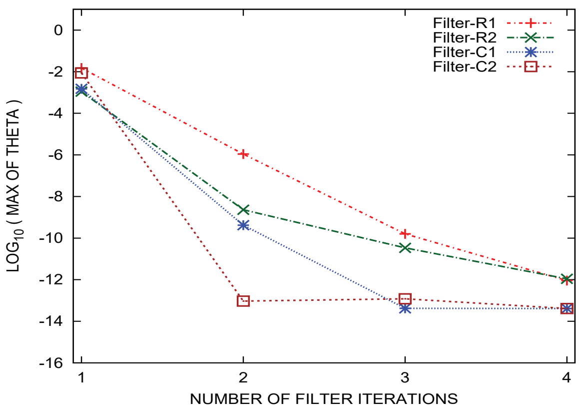

図 4 反復回数に対する最大の相対残差，各フィルタ（倍精度だけ）

6.2 倍精度と単精度を混合して用いた計算の例

次にフィルタ（R1，R2，C1，C2）を用いて，「標準精度」と「低精度」をそれぞれ IEEE754 の倍精度（2 進 64 ビット，有効 15.95 桁）と単精度（2 進 32 ビット，有効 7.22 桁）とする混合精度により計算を行った結果を示す（両者の精度の有効桁数の比は 2.21 倍である）．

近似固有対のうちで固有値が区間 $[a, b]$ にあるものの相対残差の最大値を**表 3**，**表 4**，**表 5**，**表 6** に示し，それらに対応するグラフを**図 5**，**図 6**，**図 7**，**図 8** に示す．グラフから $\ell = 3$ と $\ell = 4$ の線はまったく重なっており，残差反復は 3 回目で完了していることがわかる．またこれら 4 通りのフィルタの例から最も優れた特性をもつ C2 の場合を除けば，直交化付きのフィルタの反復を 2 回だけ行なう場合には，残差反復も $\ell = 2$ で十分であることがわかる．

またそれぞれの計算の経過時間を**表 7**，**表 8**，**表 9**，**表 10** に示す．これらの表からは，同じフィルタと反復回数 IT に対しては ℓ が 2 あるいはそれ以上の場合について，残差反復法を用いずに倍精度だけで計算した場合に比べて残差反復法を用いて混合精度により計算した方が経過時間が長いことがわかる．つまり残差反復法を用いて混合精度で計算した方が不利になっている．その理由としては，いま用いている計算機の CPU は単精度と倍精度についての演算性能の比も記憶の語転送性能の比もほぼ 2 対 1 になっており，演算主体の処理は単精度が倍精度に比べてほぼ 2 倍早い．しかし残差反復法の中では単精度で前進後退代入を 2 度以上行なうことや，残差を求めるための倍精度の計算や異なる精度間の変換の手間が掛かるので，それよりも残差反復を使用しないで倍精度だけで 1 回だけ前進後退代入を行う方が時間が掛からなかったと考えられる（ただし係数行列の分解の計算や特に分解の保持に必要なバイト単位での記憶量については，単精度は倍精度に比べて有利になる）．

7. 解を四倍精度で求めた実験の例

IEEE754 の四倍精度（2 進 128 ビット，有効 34.02 桁）でフィルタ（R1，R2，C1，C2）を IT 回反復して用いて一般固有値問題の固有対を求めた例を示す．ただし四倍精度演算は用いた CPU の機能には無くて，ソフトウェアで実現しているので計算の速度がかなり遅いので，実験で扱う問題の規模をかなり小さくして，有限要素法の要素分割を $(N_1, N_2, N_3) = (20, 30, 40)$ と倍精度の場合に比べて粗いものにした．それから導かれる一般固有値問題の係数行列 A と B は，次数が $N = 24,000$ で下帯幅は $w_L = 621$ である．

求めようとする固有対の固有値 λ の区間については，解を倍精度で求める場合と同じく，下端の固有値をフィルタ R1 と R2 を用いて求める場合については $[0, 100]$，中間の固有値をフィルタ C1 と C2 を用いて求める場合については $[100, 200]$ とした．なお，この四倍精度の場合の一般固有値問題では，固有値 λ が固有値分布の下端の区間 $[0, 100]$ にある固有対の数は 378 であり，中間の区間 $[100, 200]$ にある固有対の数は 684 である．

最初に与えるランダムなベクトルの数についても，解を

倍精度で求める場合と同じく，固有値が下端の区間にある固有対を求める場合は $m = 800$ とし，中間の区間にある固有対を求める場合は $m = 1,300$ とした．

表 3 最大の相対残差，フィルタ R1（DP-SP 混合）

IT	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
1	2.7E-01	1.5E-02	1.5E-02	1.5E-02
2	2.7E-04	1.3E-06	1.1E-06	1.2E-06
3	2.8E-04	4.3E-09	1.7E-10	1.9E-10
4	2.8E-04	4.3E-09	1.3E-13	8.6E-14

表 4 最大の相対残差，フィルタ R2（DP-SP 混合）

IT	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
1	6.7E-01	1.1E-03	1.2E-03	1.1E-03
2	2.6E-04	2.5E-09	2.3E-09	2.6E-09
3	2.6E-04	2.4E-09	6.1E-11	3.7E-11
4	2.6E-04	2.4E-09	9.6E-14	8.6E-14

表 5 最大の相対残差，フィルタ C1（DP-SP 混合）

IT	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
1	1.8E-01	1.4E-03	1.8E-03	1.4E-03
2	3.3E-05	4.2E-10	4.2E-10	4.0E-10
3	2.1E-05	4.5E-11	1.8E-14	1.8E-14
4	2.1E-05	4.5E-11	8.4E-15	8.1E-15

表 6 最大の相対残差，フィルタ C2（DP-SP 混合）

IT	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
1	1.5E+00	8.0E-03	5.5E-03	1.3E-02
2	3.2E-05	5.1E-11	8.0E-14	8.8E-14
3	2.1E-05	3.6E-11	1.5E-13	1.7E-13
4	2.3E-05	3.7E-11	7.6E-15	7.9E-15

表 7 経過時間 (秒)，フィルタ R1（DP-SP 混合）

IT	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
1	339	508	703	897
2	502	889	1,279	1,661
3	666	1,240	1,784	2,353
4	815	1,550	2,315	3,118

表 8 経過時間 (秒)，フィルタ R2（DP-SP 混合）

IT	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
1	408	687	981	1,280
2	636	1,246	1,825	2,419
3	869	1,735	2,611	3,467
4	1,084	2,237	3,386	4,505

表 9 経過時間 (秒)，フィルタ C1（DP-SP 混合）

IT	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
1	806	1,293	1,798	2,294
2	1,334	2,333	3,338	4,372
3	1,812	3,024	4,350	5,723
4	2,211	3,816	5,597	7,382

表 10 経過時間 (秒)，フィルタ C2（DP-SP 混合）

IT	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$
1	936	1,504	2,148	2,711
2	1,563	2,737	3,742	4,934
3	2,026	3,594	5,085	6,798
4	2,602	4,420	6,574	8,688

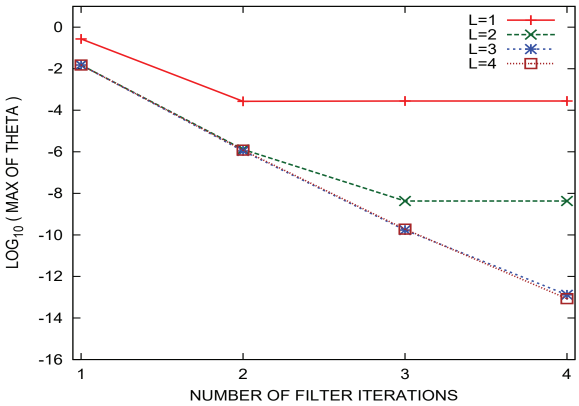

図 5 フィルタ R1：反復回数と最大の相対残差（DP-SP 混合）

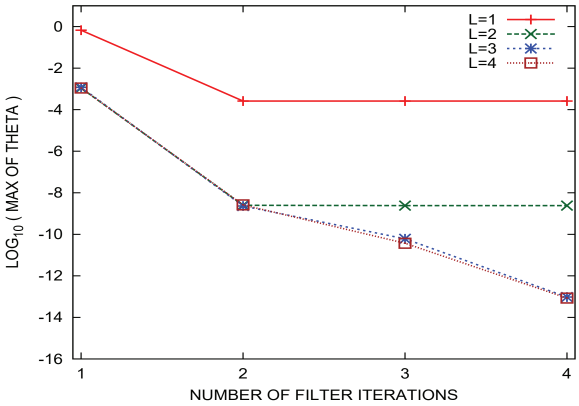

図 6 フィルタ R2：反復回数と最大の相対残差（DP-SP 混合）

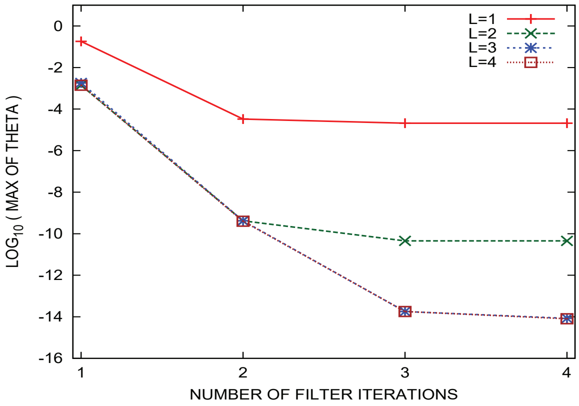

図 7 フィルタ C1：反復回数と最大の相対残差（DP-SP 混合）

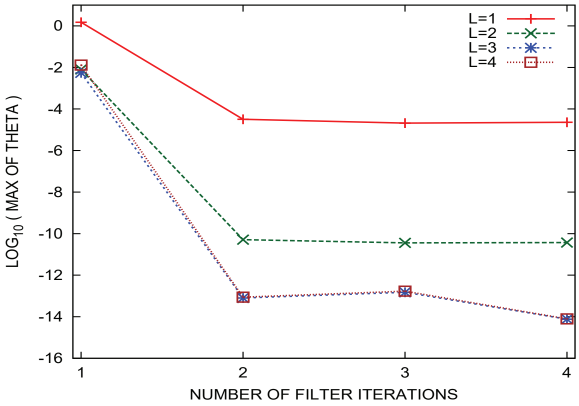

図 8 フィルタ C2：反復回数と最大の相対残差（DP-SP 混合）

7.1 四倍精度だけを用いた計算の例

まず四倍精度だけを用いて計算を行った例を示す．フィルタ（R1，R2，C1，C2）それぞれについて，フィルタと再直交化の組合わせを反復した回数 IT に対して得られた近似固有対のうちで固有値が区間 $[a, b]$ にあるものの相対残差の最大値を**表 11** に示し，それらをグラフにプロットしたものを**図 9** に示す．各フィルタについてのグラフが直線的であることからわかるように，相対残差の最大値は反復回数 IT に対して毎回一定の比率で減少している．

またそれぞれの場合の経過時間を**表 12** に示す．

表 11 反復回数と最大の相対残差，各フィルタ（四倍精度だけ）

IT	フィルタ R1	フィルタ R2	フィルタ C1	フィルタ C2
1	2.7E-03	1.5E-04	5.9E-05	6.0E-07
2	1.1E-07	1.6E-10	1.2E-11	1.8E-15
3	7.4E-12	3.5E-16	1.0E-17	1.8E-23
4	3.8E-16	7.8E-22	9.2E-24	2.2E-31

表 12 反復回数と経過時間 (秒)，各フィルタ（四倍精度だけ）

IT	フィルタ R1	フィルタ R2	フィルタ C1	フィルタ C2
1	1,565	1,906	6,552	7,617
2	2,892	3,495	12,038	14,218
3	3,649	4,545	16,360	19,305
4	4,416	5,653	20,676	24,279

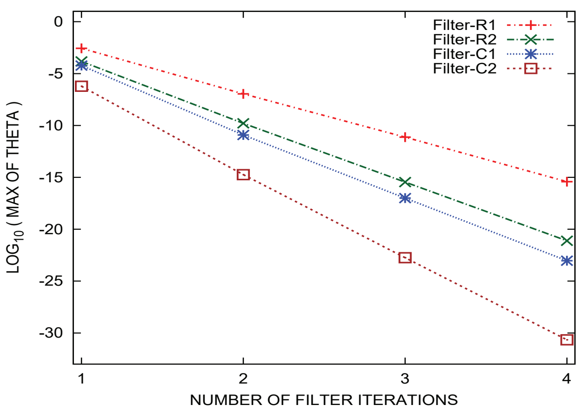

図 9 （各フィルタ）反復回数と最大の相対残差（四倍精度だけ）

7.2 四倍精度と倍精度を混合して用いた計算の例

次に，「標準精度」と「低精度」を四倍精度（128 ビット，有効 34.02 桁）と倍精度（64 ビット，有効 15.95 桁）として残差反復法の計算を混合精度により行った結果の例を示す（この場合に，2 つの精度の有効桁数の比は 2.13 である）．

フィルタ（R1，R2，C1，C2）それぞれにより得られた近似固有対で固有値が区間 $[a, b]$ にあるものの相対残差の最大値を**表 15**，**表 16**，**表 17**，**表 18** に示し，それらをグラフにプロットしたものを**図 10**，**図 11**，**図 12**，**図 13** に示す．まず，ℓ が 1 と 2 の場合の結果はフィルタの反復回数 IT が 2 以下であるときにはほぼ一致するが，最大の相対残差が小さくなってくる IT が 3 以上では違いが見られる．$\ell = 1$ の場合の最大の相対残差（各グラフ中の赤い線）は，レゾルベントの作用を倍精度だけを用いて計算した場合と同じになる．また各図において ℓ が 2，3，4 のそれぞれの場合のグラフの線（緑，青，茶）はほとんどお互いに重なっており違いが見えない（フィルタが C2 の場合にだけ，最大の相対残差が極めて小さい IT=4 のときにだけ，$\ell = 2$ の場合と ℓ が 3 以上の場合には違いが見える）．これらの計算例については，四倍精度と倍精度の混合による残差反復法の計算ではその反復回数 ℓ は 2 で十分であり，それを越えて $\ell = 3$ や $\ell = 4$ にしても無駄であると言える．

またそれぞれのフィルタを用いた場合の経過時間を**表 19**，**表 20**，**表 21**，**表 22** に示す．これらの残差反復法を用いて混合精度で計算した場合の経過時間の表と，残差反復法を用いずに四倍精度だけで計算した場合の経過時間の表（表 12）の値をフィルタの反復回数 IT が同じものについて比較すると，前者の場合が後者の場合よりも経過時間がかなり短縮できていることがわかる．たとえば比較を容易にするために前者の場合の各フィルタの反復回数 IT に対する経過時間を $\ell = 2$ についてだけ集めたものを**表 13** に示す．時間が短いことが表 12 と比べてわかる．最大の相対残差については $\ell = 2$ ではどちらもほぼ同等である．

表 13 反復回数と経過時間 (秒)，各フィルタ（QP-DP 混合，$\ell = 2$ の場合）

IT	フィルタ R1	フィルタ R2	フィルタ C1	フィルタ C2
1	967	1,083	2,585	2,943
2	1,869	2,112	4,638	5,568
3	2,149	2,462	5,514	6,328
4	2,532	2,872	6,399	7,301

7.3 四倍精度と単精度を混合して用いた計算の例

さらに「標準精度」と「低精度」を四倍精度と単精度として混合精度で残差反復法を用いて計算を行った場合について，相対残差の最大値を**表 23**，**表 24**，**表 25**，**表 26** に示し，それらをグラフにプロットしたものを**図 14**，**図 15**，**図 16**，**図 17** に示す．フィルタ C2 の場合のグラフ図 17 を見ると，四倍精度だけで計算した場合とくらべて残差反復の回数が $\ell = 4$ でもまだ最大の相対残差は同じ程度にまで達していないが，これは 2 つの精度の有効桁数の比が $34.02/7.22 = 4.71$ で 4 よりも大きいからであろう．

それぞれの経過時間を**表 27**，**表 28**，**表 29**，**表 30** に示す．それらのうちで $\ell = 4$ だけを集めたものを**表 14** に示す．四倍精度だけで計算を行った場合（表 12）よりもかなり時間が短かい（ただしこの結果は「低精度」が倍精度で $\ell = 2$ とした場合（表 13）よりも時間が長い）．

表 14 反復回数と経過時間 (秒)，各フィルタ（QP-SP 混合，$\ell = 4$ の場合）

IT	フィルタ R1	フィルタ R2	フィルタ C1	フィルタ C2
1	1,105	1,332	3,056	3,476
2	2,145	2,603	5,555	6,565
3	2,703	3,210	6,869	7,989
4	3,212	3,879	8,248	9,526

表 15 最大の相対残差，フィルタ R1（QP-DP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	2.8E-03	2.6E-03	2.6E-03	2.5E-03
2	1.1E-07	1.1E-07	1.0E-07	1.0E-07
3	7.0E-12	6.8E-12	6.9E-12	6.5E-12
4	1.3E-13	3.5E-16	3.9E-16	3.6E-16

表 16 最大の相対残差，フィルタ R2（QP-DP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	1.6E-04	1.4E-04	1.4E-04	1.4E-04
2	1.8E-10	1.8E-10	1.6E-10	1.8E-10
3	1.6E-13	3.6E-16	3.6E-16	4.4E-16
4	1.6E-13	8.9E-22	8.1E-22	8.4E-22

表 17 最大の相対残差，フィルタ C1（QP-DP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	6.1E-05	7.2E-05	4.1E-05	6.5E-05
2	1.2E-11	1.3E-11	1.3E-11	1.4E-11
3	6.8E-15	1.0E-17	1.0E-17	9.8E-18
4	7.0E-15	8.6E-24	8.6E-24	8.6E-24

表 18 最大の相対残差，フィルタ C2（QP-DP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	7.7E-07	7.3E-07	5.1E-07	5.6E-07
2	9.0E-15	1.8E-15	1.9E-15	1.7E-15
3	6.6E-15	1.7E-23	1.8E-23	1.7E-23
4	6.8E-15	9.5E-30	1.8E-31	2.0E-31

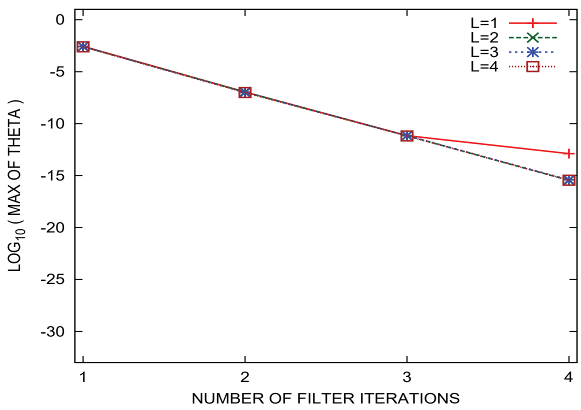

図 10 フィルタ R1：反復回数と最大の相対残差（QP-DP 混合）

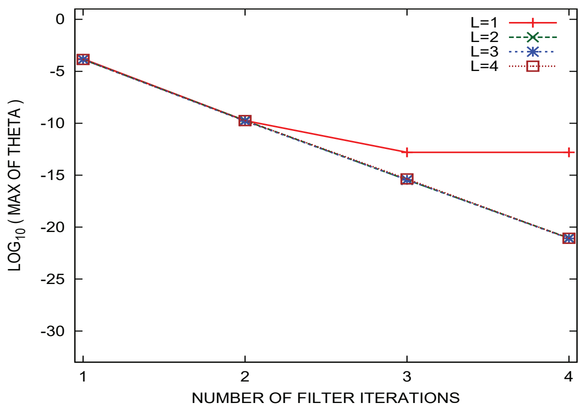

図 11 フィルタ R2：反復回数と最大の相対残差（QP-DP 混合）

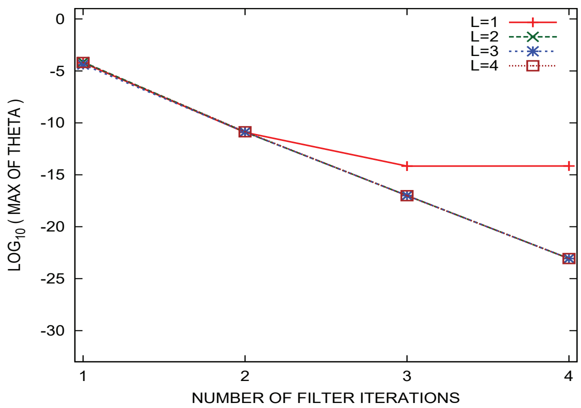

図 12 フィルタ C1：反復回数と最大の相対残差（QP-DP 混合）

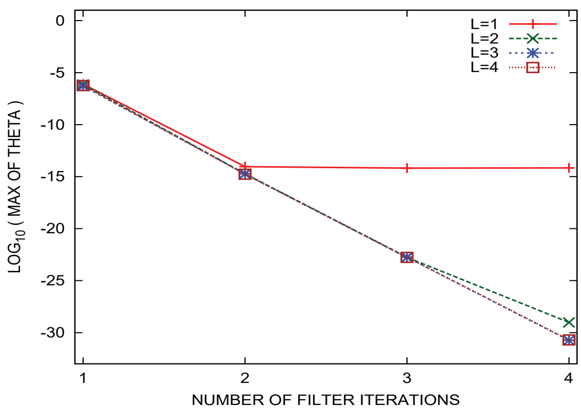

図 13 フィルタ C2：反復回数と最大の相対残差（QP-DP 混合）

表 19 経過時間 (秒)，フィルタ R1（QP-DP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	886	967	1,034	1,165
2	1,738	1,869	2,016	2,224
3	1,898	2,149	2,413	2,715
4	2,188	2,532	2,882	3,207

表 20 経過時間 (秒)，フィルタ R2（QP-DP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	963	1,083	1,238	1,341
2	1,816	2,112	2,340	2,564
3	2,122	2,462	2,870	3,239
4	2,405	2,872	3,379	3,929

表 21 経過時間 (秒)，フィルタ C1（QP-DP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	2,338	2,585	2,836	3,220
2	4,177	4,638	5,142	5,652
3	4,841	5,514	6,266	7,040
4	5,485	6,399	7,358	8,353

表 22 経過時間 (秒)，フィルタ C2（QP-DP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	2,645	2,943	3,228	3,573
2	4,903	5,568	6,045	6,666
3	5,474	6,328	7,217	8,160
4	6,214	7,301	8,374	9,698

表 23 最大の相対残差，フィルタ R1（QP-SP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	6.3E-02	2.7E-03	3.1E-03	2.7E-03
2	4.3E-05	1.1E-07	1.1E-07	1.1E-07
3	4.3E-05	9.5E-11	6.5E-12	6.1E-12
4	4.3E-05	9.4E-11	3.8E-16	3.3E-16

表 24 最大の相対残差，フィルタ R2（QP-SP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	2.6E-01	1.5E-04	1.7E-04	1.6E-04
2	4.4E-05	1.8E-10	1.6E-10	1.7E-10
3	4.5E-05	4.5E-11	3.8E-16	3.6E-16
4	4.5E-05	4.4E-11	3.1E-17	7.8E-22

表 25 最大の相対残差，フィルタ C1（QP-SP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	2.6E-02	4.9E-05	5.1E-05	6.5E-05
2	5.0E-06	1.3E-11	1.3E-11	1.2E-11
3	3.7E-06	2.8E-12	9.9E-18	1.0E-17
4	3.7E-06	2.8E-12	2.1E-18	9.1E-24

表 26 最大の相対残差，フィルタ C2（QP-SP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	8.4E-01	1.2E-06	4.7E-07	5.2E-07
2	4.8E-06	3.8E-12	1.8E-15	1.8E-15
3	3.5E-06	2.9E-12	2.2E-18	1.7E-23
4	3.6E-06	2.8E-12	2.2E-18	1.8E-24

表 27 経過時間 (秒)，フィルタ R1（QP-SP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	872	937	1,046	1,105
2	1,680	1,808	1,979	2,145
3	2,001	2,180	2,416	2,703
4	2,316	2,526	2,861	3,212

表 28 経過時間 (秒)，フィルタ R2（QP-SP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	920	1,072	1,208	1,332
2	1,801	2,051	2,313	2,603
3	2,198	2,467	2,848	3,210
4	2,549	2,877	3,353	3,879

表 29 経過時間 (秒)，フィルタ C1（QP-SP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	2,325	2,564	2,788	3,056
2	4,112	4,652	5,088	5,555
3	4,994	5,487	6,244	6,869
4	5,734	6,448	7,258	8,248

表 30 経過時間 (秒)，フィルタ C2（QP-SP 混合）

IT	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$
1	2,506	2,872	3,188	3,476
2	4,565	5,416	6,002	6,565
3	5,626	6,325	7,151	7,989
4	6,512	7,363	8,371	9,526

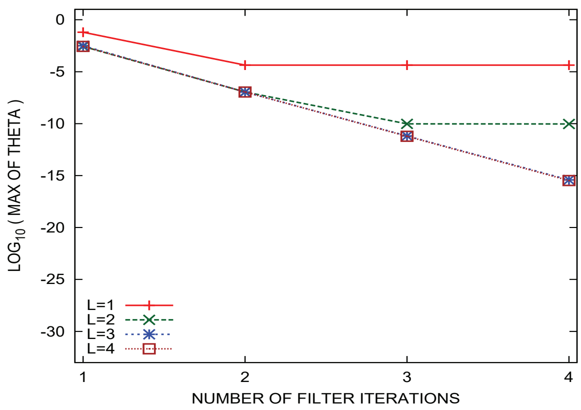

図 14 フィルタ R1：反復回数と最大の相対残差（QP-SP 混合）

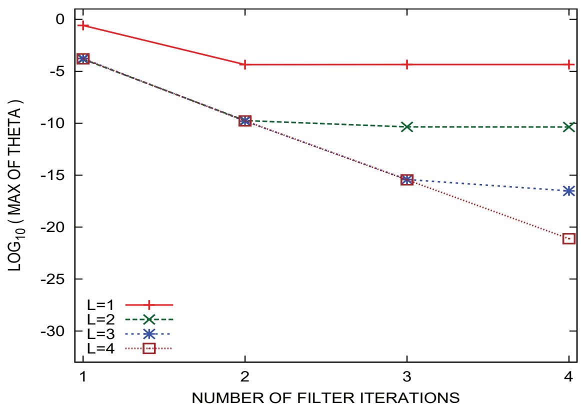

図 15 フィルタ R2：反復回数と最大の相対残差（QP-SP 混合）

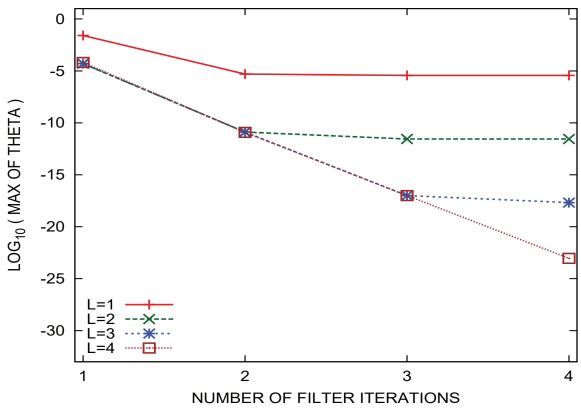

図 16 フィルタ C1：反復回数と最大の相対残差（QP-SP 混合）

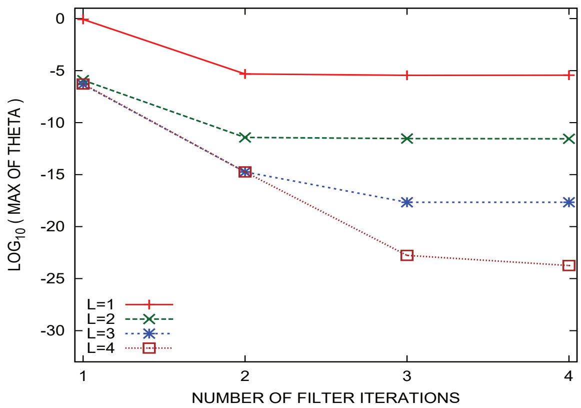

図 17 フィルタ C2：反復回数と最大の相対残差（QP-SP 混合）

8. まとめ

実対称定値一般固有値問題に対して，必要とする範囲の固有値を持つ固有ベクトルを良く通過させるフィルタを用意して，ランダムなベクトルから始めてそれに対して再直交化とフィルタの作用の組み合わせを数回適用することにより，必要な固有値を持つ固有ベクトルで張られた不変部分空間の近似空間の基底を得る．そのようにして得られた基底に対して Rayleigh-Ritz 法を適用することで，必要とする固有値を持つ固有対が一斉に得られる．

今回用いたフィルタは単一のレゾルベントの多項式であり，レゾルベントの作用を与えるための連立 1 次方程式を直接法で解く場合には，複数のレゾルベントを用いて構成されたフィルタに比べて，行列分解の手間と分解結果を保持するための記憶容量の両方が少ないという利点がある．しかし，他方で単一のレゾルベントから構成されるフィルタの伝達特性はあまり良くすることができないという難点がある．そこで，フィルタを反復して用いるが，単純に反復を行なうと必要とする固有値を持つ固有ベクトルに対する伝達率の不均一により，ベクトルに含まれている伝達率が相対的に小さい固有ベクトルの情報の精度が反復に伴って低下するので，それをなるべく防止するためにフィルタの適用ごとに再直交化を行なっている．

フィルタを適用する計算の主要部であるレゾルベントの作用は連立 1 次方程式を解くことに帰着されるが，本報告ではその解法として直接法の利用を前提としている．ただし本報告では新たに，直接法の計算を「通常精度」の数値と演算だけで行うかわりに，残差反復法を用いて計算の一部を「低精度」の数値と演算で行い，連立 1 次方程式の近似解に対して方程式の残差に対応する修正を加える処理を反復することで精度を段階的に高めて解く方法を採用してみた．これにより，行列分解や分解結果を利用した前進後退代入の計算の部分は「低精度」の数値と演算だけを用いて実行できる．また「低精度」では行列分解を保持するための記憶量が「標準精度」に比べて少ないことも利点である．

実験ではこの残差反復法を取り入れた計算を実際に行ってみた．まず「通常精度」と「低精度」をそれぞれ IEEE754 の倍精度（64 ビット）と単精度（32 ビット）にした場合には，連立 1 次方程式の解法に残差反復法を用いない場合に比べて残差反復法を用いた場合は経過時間が長くなってしまい逆効果であった．その理由はおそらく，使用した計算機の CPU では単精度での計算速度が倍精度のものに比べてほぼ 2 倍だからであろう．残差反復法の内部での繰り返しは少なくとも 2 回行うことや，さらに方程式の残差を求める計算や数値の精度変換などの余分な手間が加わるので，その結果として処理が遅くなったように思われる．

しかし，もしも使用する計算機システムの単精度での演算処理が倍精度のものに比べて 2 倍を越えてずっと速ければ，残差反復法を用いて混合精度で計算を行う方が有利になる場合もあることが期待できる．我々はこのことを以下のようにして「間接的に」示すことができた．今回の報告で用いている計算機システムは，プログラム言語からは IEEE754 の四倍精度（128 ビット）を使うことができるが，実際には倍精度や単精度の場合とは異なりハードウェア命令としては四倍精度演算が備わっていないのでソフトウェア的に複数の命令の組み合わせにより実現されている．そのため四倍精度の演算速度は倍精度や単精度に比べてかなり遅い．このことに着目して，「通常精度」を四倍精度とし，「低精度」を倍精度あるいは単精度とする実験も行ってみた．するとこれらの場合には期待したように，四倍精度だけで計算を行った場合に比べて，残差反復法を用いて混合精度で計算を行った方が経過時間をかなり短くできることが実際に確認できた．

参考文献

[1] James H. Wilkinson: *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, H. J., (1963).

[2] Heinz Rutishauser : The Jacobi method for real symmetric matrices, *Numer. Math.*,Vol.9,pp.1–10(1966). Also in book: Bauer F.L.(eds) *Linear Algebra, Handbook for Automatic Computation*,Vol.2,Springer (1971).

[3] Cleve B. Moler : Iterative refinement in floating point, *J. ACM*, Vol.15, No.2 (1967).

[4] Heinz Rutishauser : Computational aspects of F.L.Bauer's simultaneous iteration method, *Numer. Math.*, Vol.13, No.1, pp.4–13 (1969).

[5] Heinz Rutishauser : Simultaneous iteration method for symmetric matrices, *Handbook for Automatic Computation*, Springer-Verlag,pp.284–302(1971). Reprinted from *Numer.Math.*,Vol.16,pp.205–223 (1970).

[6] 村田 健郎，小国 力，唐木 幸比古 :「スーパーコンピュータ：科学技術計算への適用」，丸善 (1985). （§8.1：ベキ乗法一族，§8.3：レーリー・リッツつきの同時逆反復法，§8.5：一般固有値問題）

[7] 村田 健郎，三好 俊郎，ドンガラ，J.J.，長谷川 秀彦 :「行列計算ソフトウェア: WS，スーパーコン，並列計算機」，丸善 (1991). （§11.2：ベキ乗法一族，S11.4：レーリー-リッツ法つきの同時逆反復法，§11.8：対称行列用の一般固有値問題）

[8] 村上 弘 : 対称一般固有値問題のフィルタ作用素を用いた不変部分空間の近似構成，**情報処理学会論文誌：コンピューティングシステム** *(ACS)*, Vol.4,No.4(ACS35),pp.1–14 (2011).

[9] Gene H. Golub and Charles F. Van Loan: *Matrix Computations*,4th Ed.,The John Hopkins Univ. Press (2013). (§8.2.4:'Orthogonal Iteration').

[10] 村上 弘：単一のレゾルベントのチェビシェフ多項式による実対称定値一般固有値問題の解法用の簡易型フィルタ，**情報処理学会論文誌：コンピューティングシステム** *(ACS)*, Vol.12,No.2(ACS64),pp.1–26 (2019).

[11] 村上 弘：直交化付きフィルタ適用による固有値問題の近似対の反復改良について，**情報処理学会研究報告**, Vol.2019-HPC-169,No.1,pp.1–31 (2019).

[12] 村上 弘：フィルタの反復適用による実対称定値一般固有値問題の近似対の改良，**情報処理学会論文誌：コンピューティングシステム** *(ACS)*, Vol.12,No.3(ACS65),pp.14–33 (2019).

[13] Hiroshi Murakami: Single-Precision Calculation of Iterative Refinement of Eigenpairs of a Real Symmetric-Definite Generalized Eigenproblem by Using a Filter Composed of a Single Resolvent, **情報処理学会研究報告**, Vol.2020-HPC-176,No.5,pp.1–9 (2020).

付　　録

A.1 実験に用いたフィルタの設計法

単一のレゾルベントを用いた簡易型のフィルタの設計法を示す．

フィルタはパラメタの 3 つ組（n，μ，$g_{\rm s}$ ）で指定するものとする（ほかのやり方として（μ，$g_{\rm s}$，$g_{\rm p}$）を指定してそれらを近似的に満たすようにすることも可能である）．ここで n はチェビシェフ多項式の次数であり，μ は遷移域と阻止域の境界位置（の正規化座標）を表す．そうして $g_{\rm s}$ は阻止域での伝達関数の大きさの上限値であり，通過域での伝達関数の値は $g_{\rm p}$ 以上 1 以下である．

フィルタを適用すると各ベクトルに含まれる固有ベクトルのうちで必要なものに対する不要なものの割合は（数値丸め誤差の影響を無視する近似のもとでは）フィルタの適用ごとに $g_{\rm s}/g_{\rm p}$ 倍以下に減少する．

以下に，フィルタの具体的な構成法を示す．ただし一般固有値問題 $A\mathbf{v} = \lambda B\mathbf{v}$ に対する，シフトが複素数 z のレゾルベントは $\mathcal{R}(z) \equiv (A - zB)^{-1}B$ である．

A.1.1 レゾルベントのシフトを実数にする場合

レゾルベントのシフトを実数にする場合には，区間 $[a, b]$ は固有値分布の下端であって，a は最小固有値 $\lambda_{\rm min}$ 以下であることが必要である．

指定されたパラメタの組（n，μ，$g_{\rm s}$）からシフト ρ とレゾルベントの係数 γ（および $g_{\rm p}$）を以下の式 (A.1) で計算する．

$$\left\{\begin{array}{lcl} \sigma & \leftarrow & \mu/\sinh\left(\frac{1}{2n}\cosh^{-1}\frac{1}{g_{\rm s}}\right), \\ \rho & \leftarrow & a-(b-a)\,\sigma, \\ \gamma & \leftarrow & (b-a)(\sigma+\mu), \\ g_{\rm p} & \leftarrow & g_{\rm s}\cosh\left\{2n\sinh^{-1}\sqrt{(\mu-1)/(1+\sigma)}\right\}. \end{array}\right. \tag{A.1}$$

すると，フィルタ $\mathcal{F}$ はシフトが実数 ρ のレゾルベントを用いて以下の式 (A.2) で与えられる．

$$\mathcal{F} \equiv g_{\rm s}\,T_n\left(2\gamma\,\mathcal{R}(\rho) - I\right). \tag{A.2}$$

$\lambda \in [a, b]$ を $t \in [0, 1]$ に移す線形変換 $t \equiv \dfrac{\lambda - a}{b - a}$ により，λ に対する正規化座標 t を定義する．引数 t の伝達関数 $g(t)$ と，引数 λ の伝達関数 $f(\lambda)$ は以下の式 (A.3) になる：

$$\left\{\begin{array}{lcl} g(t) & = & g_{\rm s}\,T_n\left(2\,\dfrac{\mu+\sigma}{t+\sigma}-1\right), \\ f(\lambda) & = & g_{\rm s}\,T_n\left(2\gamma\,\dfrac{1}{\lambda-\rho}-1\right). \end{array}\right. \tag{A.3}$$

A.1.2 レゾルベントのシフトを虚数にする場合

レゾルベントのシフトを虚数にする場合は，区間 $[a, b]$ の位置は任意に設定できる．指定されたパラメタの組（n，μ，$g_{\rm s}$）から式 (A.4) でシフト ρ' とレゾルベントの係数 γ'（および $g_{\rm p}$）を計算する．

$$\left\{\begin{array}{lcl} \sigma & \leftarrow & \mu/\sinh\left(\frac{1}{2n}\cosh^{-1}\frac{1}{g_{\rm s}}\right), \\ \rho' & \leftarrow & \frac{a+b}{2}+\left(\frac{b-a}{2}\right)\sigma\sqrt{-1}, \\ \gamma' & \leftarrow & \left(\frac{b-a}{2}\right)\frac{\mu^2+\sigma^2}{\sigma}, \\ g_{\rm p} & \leftarrow & g_{\rm s}\cosh\left\{2n\sinh^{-1}\sqrt{(\mu^2-1)/(1+\sigma^2)}\right\}. \end{array}\right. \tag{A.4}$$

すると，フィルタ $\mathcal{F}$ はシフトが虚数 ρ' のレゾルベントを用いて以下の式 (A.5) で与えられる．

$$\mathcal{F} \equiv g_{\rm s}\,T_n\left(2\gamma'\,{\rm Im}\,\mathcal{R}(\rho') - I\right). \tag{A.5}$$

$\lambda \in [a, b]$ を $t \in [-1, 1]$ に移す線形変換 $\lambda \equiv \frac{1}{2}(a+b) + \frac{1}{2}(b-a)t$ で，λ に対する正規化座標 t を定義する．

引数 t の伝達関数 $g(t)$ と引数 λ の伝達関数 $f(\lambda)$ は以下の式 (A.6) になる．

$$\left\{\begin{array}{lcl} g(t) & = & g_{\rm s}\,T_n\left(2\,\dfrac{\mu^2+\sigma^2}{t^2+\sigma^2}-1\right), \\ f(\lambda) & = & g_{\rm s}\,T_n\left(2\,\gamma'\,{\rm Im}\,\dfrac{1}{\lambda-\rho'}-1\right). \end{array}\right. \tag{A.6}$$

A.2 例題に用いた実対称定値一般固有値問題

例題とした実対称定値一般固有値問題 (A.7) は，1 辺の長さ π の立方体領域において零ディリクレ境界条件を課された 3 次元ラプラス作用素の固有値方程式 (A.8) を有限要素法（FEM）により離散化近似して得られたものである．

$$A\,\mathbf{v} = \lambda B\,\mathbf{v}\,. \tag{A.7}$$

$$-\Delta\Psi(x, y, z) = \lambda\Psi(x, y, z)\,. \tag{A.8}$$

立方体の各辺方向を N_1+1, N_2+1, N_3+1 に等分割して，区間の直積による直方体を FEM の要素とした（**図 A·1**）．各要素内の基底関数には各辺方向の 3 重線形関数を用いた．一般固有値問題の行列 A と B の次数は $N = N_1N_2N_3$ となる．

いま $N_1 \le N_2 \le N_3$ であるとして，帯幅が小さくなるよ

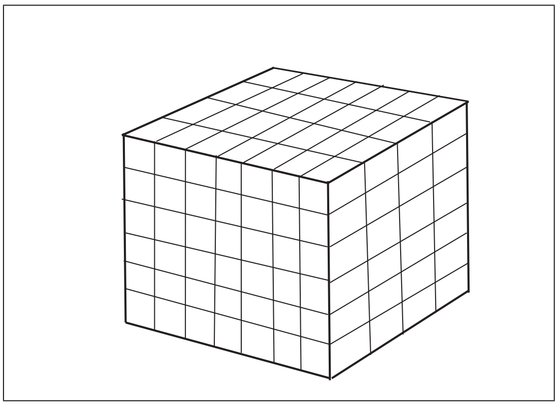

図 A·1　FEM 要素分割の概念図．$(N_1, N_2, N_3) = (3, 5, 6)$ の例．

うに基底関数に番号を付けると，下帯幅（対角要素を含めない）は $w_L = 1 + N_1 + N_1 N_2$ にできる．離散化された一般固有値問題に対してフィルタ対角化法を適用して，固有値が区間 $[a, b]$ にある固有対の近似を求める．この例題の厳密な固有値は簡単な数式で計算できるので，任意区間内にある固有値の数は，厳密値の数え上げにより求められる．

A.2.1 例題の固有値の厳密値を与える式

辺長 π の立方体の各辺方向をそれぞれ N_1+1，N_2+1，N_3+1 に等分割した場合には，例題の 3 次元問題の固有値は添字の組 (k_1, k_2, k_3) で識別されて，各方向ごとの 1 次元問題の固有値の和として式 (A.9) で表される．そうして各 1 次元問題の固有値は式 (A.10) で与えられる．ただしここで $\theta_k \equiv \frac{\pi k}{N+1}$ である．

$$E^{[N_1,\,N_2,\,N_3]}_{(k_1,\,k_2,\,k_3)} = \mathcal{E}^{[N_1]}_{k_1} + \mathcal{E}^{[N_2]}_{k_2} + \mathcal{E}^{[N_3]}_{k_3}\ ,\ k_i{=}1, 2, \ldots, N_i \tag{A.9}$$

$$\mathcal{E}^{[N]}_{k} = \frac{6\,k^2\,(\sin\theta_k/\theta_k)^2}{(1+\cos\theta_k)\,(2+\cos\theta_k)},\ k{=}1, 2, \ldots, N\,. \tag{A.10}$$

A.3 近似固有対の相対残差

近似固有対 $(\lambda, \mathbf{v})$ の相対残差 Θ を式 (A.11) で定義する．

$$\Theta \ \equiv\ \frac{||\,A\,\mathbf{v} - \lambda B\,\mathbf{v}\,||}{||\,\lambda B\,\mathbf{v}\,||} = \frac{||\,\mathbf{r}\,||}{||\,\lambda B\mathbf{v}\,||}\,. \tag{A.11}$$

Θ の値はベクトル $\mathbf{v}$ の規格化には依らず，行列 A と B に共通の尺度を乗じても不変である．今回の実験にはベクトルのノルム $||\cdot||$ として 2-ノルムを用いた．この相対残差 Θ の値が小さいほど近似固有対の品質は良いと評価する．相対残差の計算では複数の近似固有対の列ベクトルをまとめて行列 V にすると，AV と BV を作るための行列 A と B 全体への記憶参照は 1 回ずつになり，しかも A と B が疎であるほど手間が減って計算が容易になる．

ベクトルのノルムとして 2-ノルムを用いた場合の Θ の幾何学的な意味は**図 A·2** に示すように，ベクトル $A\mathbf{v}$ と $\lambda B\mathbf{v}$ の挟む角を ϕ とすると，関係式 $\sin\phi \le ||\mathbf{r}||/||\lambda B\mathbf{v}|| = \Theta$ が成り立つことである．

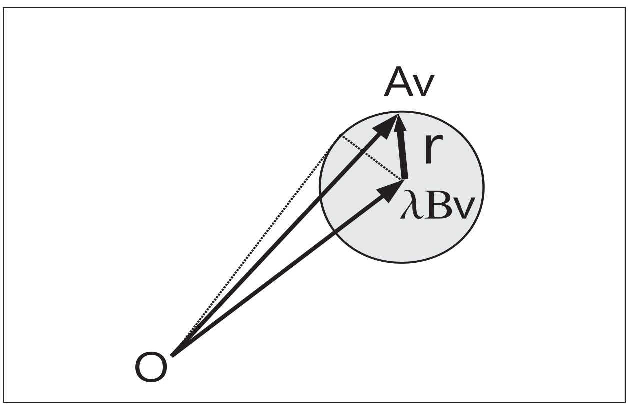

図 A·2 相対残差 Θ の幾何学的意味

A.4 再直交化付きフィルタ反復の方法の概要

フィルタの適用を正規直交化と組み合わせて反復することにより改良された近似固有対を求める計算手順を**図 A·3** に示す．今回の実験では，同一のフィルタを IT 回（1~ 4 回）適用した．そうして毎回のフィルタの適用ごとに必ず B-正規直交化を施している．B-正規直交化は閾値付きの B-特異値分解（B-SVD）を用いて行った．その閾値は使用した「標準精度」に対するマシンイプシロンの 100 倍に設定し，特異値が閾値以下の特異ベクトルは切断した．なおフィルタの準備として行列分解を行なうシフト行列は，計算を混合精度で行なう場合には「低精度」のものである．

(1) フィルタ $\mathcal{F}$ の準備として，ここでシフト行列を分解する；
(2) $Y^{(0)} \leftarrow$ ランダムなベクトル m 個の組；
(3) for $i := 1, 2, \ldots,$ `IT` do
　　$X^{(i)} \leftarrow Y^{(i-1)}$ の B-正規直交化；
　　$Y^{(i)} \leftarrow \mathcal{F}\,X^{(i)}$；
　enddo
　　注：途中の B-正規直交化においてベクトルの組 $Y^{(i-1)}$ の階数不足により $X^{(i)}$ の階数が切断を受けて低下したら，その後の $X^{(i)}$ や $Y^{(i)}$ のベクトルの数はその低下した階数へ変更．
(4) 上記のステップ (3) で得られた最後の X と Y について，以下のようにする．
　「必要な固有値全部を持つ不変部分空間」の近似空間の基底 Z を Y の列の線形結合で構成する（その際に X および フィルタの伝達特性の値 $g_{\rm s}$ と $g_{\rm p}$ の情報も用いる）[8].
(5) 基底 Z に Rayleigh-Ritz 法を適用して得られた Ritz 対を，元の一般固有値問題の近似対とする．

図 A·3 正規直交化付きフィルタ適用の反復による対角化の手順

A.4.1 フィルタ適用による固有ベクトルの含有比の変化

いま対象としている一般固有値問題の固有対のすべてを $(\lambda_i, \mathbf{v}_i)$，$i{=}1, 2, \ldots, N$ とする．任意のベクトル $\mathbf{x}$ にフィルタ $\mathcal{F}$ を適用すると，対応する固有ベクトル展開では，任意の i 番目の固有ベクトル $\mathbf{v}_i$ の係数はそれにさらにフィルタの伝達率 $f(\lambda_i)$ を乗じたものになる．今回のフィルタの伝達関数の設定では固有値 λ が通過域にあれば $g_{\rm p} \le f(\lambda) \le 1$ であり，固有値 λ が阻止域にあれば $|f(\lambda)| \le g_{\rm s}\ (\ll g_{\rm p})$ である．よってフィルタをベクトル $\mathbf{x}$ に適用すると，その固有ベクトル展開では固有値 λ_j が通過域にある固有ベクトル $\mathbf{v}_j$ の係数はどれもその $f(\lambda_j)$ 倍に変わり，固有値が阻止域にある任意の固有ベクトルに対する展開係数の大きさは $g_{\rm s}$ 以下の数が乗じられたものに変わる．

よって「（多くの）阻止したい固有ベクトルの強度」の「通過させたい j 番目の固有ベクトルの強度」に対する割合は，フィルタを適用すると（どれも）適用する前の $g_{\rm s}/f(\lambda_j)$ 倍以下になる．この比の値は 1 よりも小さく，それは「特定の信号 j の強度に対するノイズの強度の比率」がフィルタの適用により受けた低減の倍率である．この倍率が小さいほど，フィルタの適用により除去したい固有値が阻止域

にある固有ベクトルの「相対的」な含有率が減ることになる．通過させたい固有ベクトルの伝達率が大きいほどこの倍率は小さくなり，フィルタの適用により相対比で考えたときの品質改良の程度が高い．

固有値が通過域にある伝達率が g_{p} 以上の固有ベクトルをすべて求める場合には，上記の比の値の上限は $g_{\mathrm{s}}/g_{\mathrm{p}}$ になる．この上限は通過域にある固有値に対する伝達率の最小値が閾値 g_{p} に等しい場合にだけ達成されるものなので，最小値がそれよりも大きければそれだけ，実際の改良は上限から予想されるよりも良くなる．

訂正（一部の数式の修正）

HPC-177 の予稿の投稿後に，不注意により数式の 1 つに誤りを入れていたことが判明いたしました．お詫びしてここにその訂正について記述させていただきます．

- 付録の ” A.1.1 レゾルベントのシフトを実数にする場合” の数式 (A.1) において，

$$\sigma \leftarrow \mu \Big/ \sinh\left(\frac{1}{2n}\,\cosh^{-1}\,\frac{1}{g_{\mathrm{s}}}\right)$$

  とあるのは誤りで，sinh の肩の指数 2 が抜けており，正しくは以下のとおりです．

$$\sigma \leftarrow \mu \Big/ \sinh^{2}\left(\frac{1}{2n}\,\cosh^{-1}\,\frac{1}{g_{\mathrm{s}}}\right)$$

計算に用いたプログラムは正しい式のものになっており，実験に於ける計算結果には影響はありません．

以上です．