

実数シフトのレゾルベントを組み合わせたフィルタによる 実対称定値一般固有値問題の下端付近の固有値を持つ固有対 の解法

村上 弘^{1,a)}

概要: 実対称定値一般固有値問題 $A\mathbf{v} = \lambda B\mathbf{v}$ (A, B が実対称行列で B は正定値) の固有対をフィルタ対角化法を用いて解く。フィルタにはレゾルベントの線形結合を用いる。本論文では固有値が固有値分布の下端付近にある固有対だけを求める場合には、フィルタを構成するレゾルベントのシフト量をすべて実数にできて、各レゾルベントの作用は実対称正定値の行列を係数とする線形方程式を解いて実現できることを示す。この方法で構成されたフィルタを用いた数値実験の例を幾つか示す。

キーワード: レゾルベント, 固有値問題, フィルタ対角化

Computing the Smallest Eigenvalues and Eigenvectors of a Real Symmetric Definite Generalized Eigenproblem by Using a Filter which is a Combination of Resolvents whose Shifts are Reals

MURAKAMI HIROSHI^{1,a)}

Abstract: We compute eigenpairs of a real symmetric definite generalized eigenproblem $A\mathbf{v} = \lambda B\mathbf{v}$ (here A and B are real symmetric matrices, and B is positive definite) by using a filter diagonalization method. For the filter, a linear combination of resolvents is used. In this report, we show that when only those eigenpairs are solved whose values are smallest in the eigenvalue distribution, all shifts of resolvents for the filter can be set to real numbers, and each application of resolvent is realized by solving a linear equation whose coefficient matrix is real and symmetric positive definite. Some examples of numerical experiments are shown that use the filters constructed by this method.

Keywords: resolvent, eigenproblem, filter diagonalization

1. はじめに

これまで行ってきた研究 [2], [3], [4], [7] により、レゾルベントの線形結合をフィルタに用いるフィルタ対角化法では、フィルタへの入力ベクトルの組とフィルタからの出力ベクトルの組を併せてフィルタの伝達特性に基づいて分析

することで、「必要な固有値を含む区間と対応する不変部分空間」の良い近似基底を構成する方法が得られている。その方法を用いれば、フィルタの遮断特性が鋭くない場合にも、(濾過するベクトルの個数を十分多くとれば) 必要な固有対の比較的良い近似が安定に得られる。

いま、実対称定値一般固有値問題で、固有値分布の下端付近の固有値を持つ比較的少数の固有対だけを解くという状況を想定する。これは有限要素法による振動モードの解析などでは普通に出てくる状況である。レゾルベントのシ

¹ 首都大学東京 数理情報科学専攻
Department of Mathematics and Information Sciences,
Tokyo Metropolitan University

^{a)} mrkmhrsh@tmu.ac.jp

フト量に実数を用いる場合は、有界な作用素となることを固有値の分布によらずに保証するためには、シフト量を最小固有値よりも小さい値に制限する必要がある。その場合に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実対称で正定値になる。連立1次方程式が正定値であるとそれを解く上では数値安定性の面で有利になる。しかし、シフト量が最小固有値より小さい実数に制限されたレゾルベントだけを用いて構成されたフィルタは遮断特性を急峻にはできない点で不利である。そこで本研究では、フィルタ対角化法にそのように構成された遮断特性が急峻でないフィルタを用いた場合に、必要な近似対がどの程度の精度で求められるかについて調査を行った。

2. レゾルベントの線形結合をフィルタに用いた対角化法の概要

対象とするのは係数行列 A, B が共に実対称で B は正定値の実対称定値一般固有値問題 $A\mathbf{v} = \lambda B\mathbf{v}$ である。今回用いたフィルタ対角化法 [3], [7] の概要は以下のようになる。

- (1) レゾルベントの線形結合であるフィルタ作用素 $F = c_\infty I + \sum_p \gamma_p \mathcal{R}(\tau_p)$ (但し、シフト τ のレゾルベントを $\mathcal{R}(\tau) \equiv (A - \tau B)^{-1} B$ と定義する) をうまく構成して、区間 $[a, b]$ から離れた固有値の固有ベクトルを強く減衰させるようにする。
- (2) 乱数で生成した縦ベクトル m 個の組を B -正規直交化した組を $X = \{X^{(1)}, X^{(2)}, \dots, X^{(m)}\}$ ($X^T B X = I$) とする。フィルタ F をこの組 X に作用させてベクトルの組 $Y = F X$ を得る。
- (3) フィルタの入力 X と出力 Y の情報に加え、フィルタ F の伝達特性も参照して、「固有値が $[a, b]$ の近傍にある不変部分空間」を良く近似する部分空間の基底 Z を $\text{span}(Y)$ 内の B -正規直交基底として作成する。
- (4) 得られた不変部分空間の近似 $\text{span}(Z)$ から固有値問題の固有対の近似を Rayleigh-Ritz 法 (subspace 法) を用いて Ritz 対として取り出す。

2.1 フィルタ作用素とその伝達関数

N 次の実対称定値一般固有値問題 $A\mathbf{v} = \lambda B\mathbf{v}$ (A, B は実対称で B は正定値) に対して、フィルタ F をレゾルベントの線形結合 $F = c_\infty I + \sum_p \gamma_p \mathcal{R}(\tau_p)$ とする。ただし、シフト τ のレゾルベントは $\mathcal{R}(\tau) \equiv (A - \tau B)^{-1} B$ とする。そのとき、固有値が λ の固有ベクトル \mathbf{v} に対しては $F\mathbf{v} = f(\lambda) \cdot \mathbf{v}$, $f(\lambda) = c_\infty + \sum_p \frac{\gamma_p}{\lambda - \tau_p}$ が成立する。この有理関数 $f(\lambda)$ は伝達関数と呼ばれ、「固有値 λ を持つ固有ベクトル」の F による伝達率を与える。逆に上記の形式を持つ有理関数 $f(\lambda)$ を与えると、それを伝達関数を持つ作用素 F をレゾルベントの線形結合で実現できる。フィルタの設計に於いては、伝達関数 $f(\lambda)$ の形状に対する制約の

組を与えて、それを満たすようにレゾルベントの個数、実係数 c_∞ 、一般には複素数であるシフト量 τ_p と係数 γ_p の組を決定する (フィルタ作用素の中の恒等演算子の項は係数 c_∞ の値が極めて小さい場合には省くことがある)。

3. 中間固有値に対するフィルタの設計法概略

「固有値分布の内部の区間 $[a, b]$ に固有値がある固有対」を求めるのにレゾルベントのシフト量として虚数を用いる場合のフィルタの設計法 [1], [7] の概略を以下に述べる。

いま $\lambda \in [a, b]$ から $t \in [-1, 1]$ への線形変換で定義される正規化座標 t を用いて (図 1), 引数 t の伝達関数 $g(t) \equiv f(\lambda)$ を定義する (図 2)。また、 $g(t)$ の最大値は 1 に規格化されているとする。いま $\mu > 1, 1 > g_{\text{pass}} \gg g_{\text{stop}} > 0$ を満たす 3 つ組 $(\mu, g_{\text{pass}}, g_{\text{stop}})$ を与えて、伝達関数 $g(t)$ の形状への制約を課す: 通過域 (passband) $|t| \leq 1$ に於いては $g_{\text{pass}} \leq g(t)$ であり、逆に $g_{\text{pass}} \leq g(t)$ となるのは通過域 $|t| \leq 1$ に於いてだけであり、阻止域 (stopbands) $|t| \geq \mu$ に於いては常に $|g(t)| \leq g_{\text{stop}}$ となるようにする。途中の $1 < |t| < \mu$ を遷移域 (transitionbands) と呼ぶ。

四種類の典型フィルタ [1] (Butterworth, Chebyshev, inverse-Chebyshev, elliptic) ではレゾルベントのシフトの個数を $2n$ (互いに複素共役な虚数の対が n 個) とするとき、三つ組 $(\mu, g_{\text{pass}}, g_{\text{stop}})$ についての制約条件が指定されると、それを満たすことのできる n の最小値 n_{min} が決まる。そうして n_{min} 以上である整数 n を指定すれば $g(t)$ が完全に決まり、 $g(t)$ の部分分数展開に対応するレゾルベントの線形結合としてフィルタ F が決まる。あるいは、制約条件を指定する三つ組が $(n, g_{\text{pass}}, g_{\text{stop}})$ であれば、それを満たすことのできる μ の最小値 μ_{min} が決まり、 μ_{min} 以上である実数 μ を指定すれば $g(t)$ が完全に決まり、 $g(t)$ の部分分数展開からフィルタ F が決まる。

例として elliptic フィルタで ($n=8, g_{\text{pass}}=0.5, g_{\text{stop}}=10^{-14}, \mu=2.34$) の (但し、係数 c_∞ を省いた) 場合の、正規化座標 t での伝達関数 $g(t)$ の極の分布 (図 3) および伝達関数の大きさ $|g(t)|$ の対数値をプロットしたグラフ (図 4) を示す。

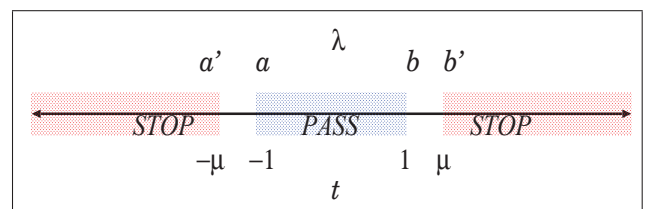


図 1 固有値分布とフィルタの設定 (中間固有値を求める場合)

4. 下端付近の固有値に対するフィルタの設計法

通過域でのフィルタ作用素 F の伝達関数の値の最大最

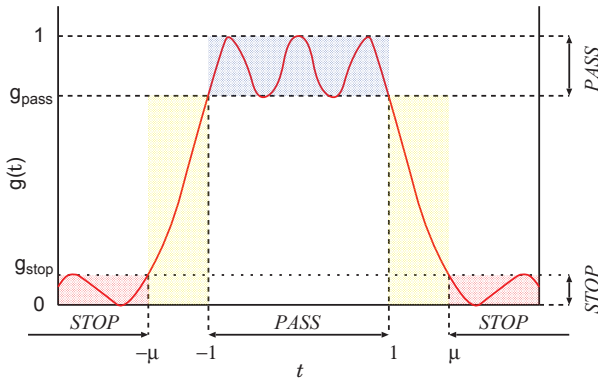


図 2 フィルタの伝達関数 $g(t)$ の概形図 (中間固有値を求める場合)

小の比は小さいことが計算精度の面から望ましい。まず \mathcal{F} が有界になるためには、その伝達関数 $f(\lambda)$ が固有値と一

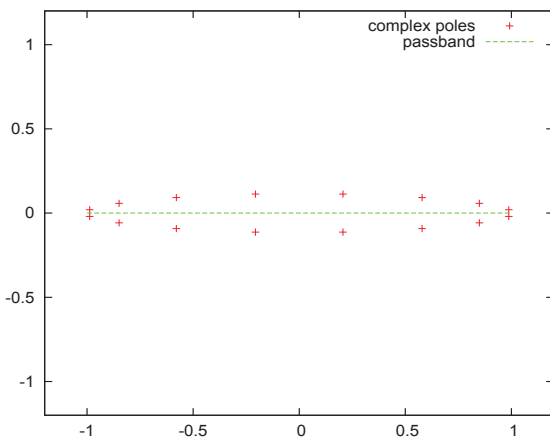


図 3 伝達関数の複素平面上の極の分布の例: elliptic フィルタ ($n=8$, $g_{\text{pass}}=0.5$, $g_{\text{stop}}=10^{-14}$, $\mu=2.34$)

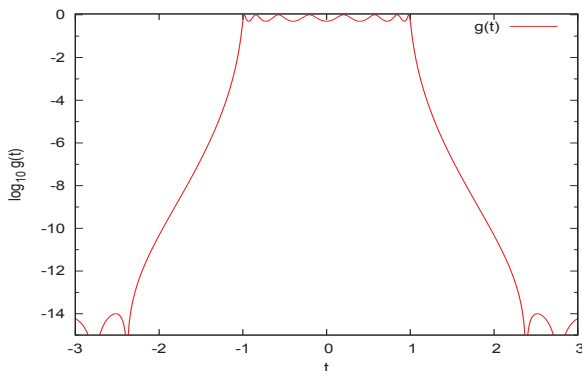


図 4 伝達関数の例: elliptic フィルタ (対数値) ($n=8$, $g_{\text{pass}}=0.5$, $g_{\text{stop}}=10^{-14}$, $\mu=2.34$)

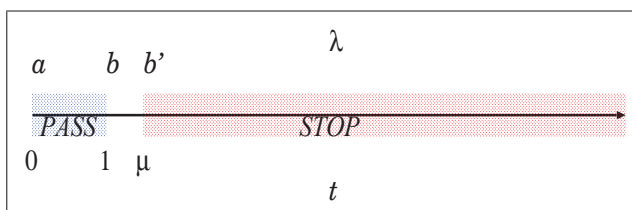


図 5 固有値分布とフィルタの設定 (最小側固有値)

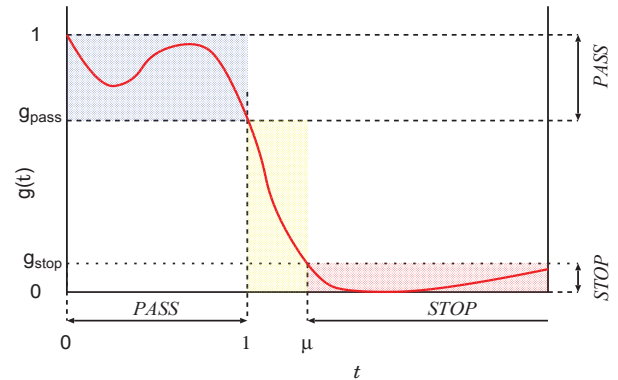


図 6 フィルタの伝達関数 $g(t)$ の概形図 (最小側固有値に対応する)

致する極を持たないことが必要である。実対称定値一般固有値問題の固有値は実数なので、 $f(\lambda)$ のすべての極を虚数に設定すれば固有値分布には依らずに作用素 \mathcal{F} の有界性が保証できるが、 $f(\lambda)$ に実数の極がある場合もそれに一致する (もしくは極端に接近する) 固有値が存在しなければ \mathcal{F} は有界である。たとえば $f(\lambda)$ のすべての極が最小固有値より小さい実数ならば、作用素 \mathcal{F} は有界である。

いま、実数 a が固有値分布の下限である (最小固有値以下である値) とする。 (たとえば定値一般固有値問題 $A\mathbf{v} = \lambda B\mathbf{v}$ の係数行列 B が正定値で A は半正定値であるならば固有値は非負で、 0 を a の値として採用できる)。固有値分布の下端付近の区間 $[a, b]$ を通過域とするフィルタを構成するには、まず $\lambda \in [a, b]$ を 1 次変換 $\lambda = \mathcal{L}(t) = a + (b-a)t$ を用いて正規化座標 $t \in [0, 1]$ に対応させる。フィルタ \mathcal{F} の伝達関数 $f(\lambda)$ のすべての極が a より小さい実数ならば、この対応で正規化座標 t の伝達関数 $g(t)$ の極はすべて負の実数になる。そこで、負数である $g(t)$ のすべての極を $-a_p$, $p=1, 2, \dots, n$ とし、 $g(t) = c_\infty + \sum_{p=1}^n \frac{c_p}{t+a_p}$ であるとするとき、それが 1 次変換で $f(\lambda) = c_\infty + \sum_{k=1}^p \frac{\gamma_p}{\lambda - \tau_p}$ に対応するならば、双方のパラメタの間には $\gamma_p/c_p = \mathcal{L}' = (b-a)$, $\tau_p = \mathcal{L}(-a_p) = a - (b-a)a_p$ の関係がある (ここで \mathcal{L}' は $\mathcal{L}(t)$ の導関数の意味である)。そうしてフィルタはレゾルベントの線形結合 $\mathcal{F} = c_\infty I + \sum_{p=1}^n \gamma_p \mathcal{R}(\tau_p)$ となる。

4.1 今回の実数シフトのフィルタの構成法

いま $t \in [0, 1]$ を通過域とするフィルタを、シフト量が負の実数であるレゾルベントの線形結合で構成する。そのために、 $\mu (> 1)$ の値を決めて、すべての極が負の実数である実有理関数 $g(t)$ をうまく構成して、 $g(t)$ の値が阻止域 $[\mu, \infty]$ では極めて微小で 0 に近く、通過域 $[0, 1]$ では 1 付近となるようにする。フィルタがレゾルベントの線形結合であれば伝達関数 $g(t)$ の極はすべて単純である。さらに簡単のために定数項 c_∞ を省略して $g(t) = \sum_{p=1}^n \frac{c_p}{t+a_p}$ であるとする。この n 個のシフト量 (レゾルベントのシフト量 τ_p に 1 次変換で対応する値) である負数 $-a_p$ と実係数 c_p , $p=1, 2, \dots, n$ とをうまく調節することで $g(t)$ が通過域

[0, 1] では 1 付近の値をとり, [0, 1] から離れると 0 に近い極めて微小な値をとるようにする. (但し, シフト量の最適あるいは望ましい配置については現在まだ決定できていない.) 最適なシフト量 $-a_p$ の組の配置の決定は一般には多次元非線形最適化問題となる. 今回は, あらかじめ個数 n と良い性質を持つシフト量の組 $\{-a_p\}$ が与えられたときに, 線形結合の係数 c_p を決める. 以下に, 今回の実験に用いた係数 c_p の決定方法を述べる.

4.1.1 区間 $[\nu, \infty)$ での最小 2 乗法による構成

ここで区間 $[\nu, \infty)$ はフィルタの阻止域 (の近似) になることを意図している.

近似誤差の 2 乗の積分値を最小化する「最小 2 乗法」を採用する場合の利点は, 連立 1 次方程式を解いて係数 c_p が求められることである.

いま ν を 1 よりも少し大きい数として, $g(t) = \sum_{p=1}^n \frac{c_p}{t+a_p}$ の 2 乗に非負の重み関数 $w^{[\text{stop}]}(t)$ を掛けた値の区間 $[\nu, \infty)$ に於ける積分値を最小化するように係数 c_p を決める.

注: 区間の端 ν の値は希望するフィルタ特性 μ と同じ値 (あるいは近い値) に設定する. そうしてフィルタの μ の値は最適化により得られた $g(t)$ の形状から g_{stop} の設定にも関係して決まるものなので, 最適化のために設定した積分区間の端の値 ν と等しくならないことがある.

但し「minimize $J^{[\text{stop}]}$, subject to $g(0)=1$ 」を Lagrange の未定乗数法で解くために, Lagrangian を $J = J^{[\text{stop}]} + \lambda^{(L)} \left(1 - \sum_{p=1}^n \frac{c_p}{a_p}\right)$ とする. ここで, $J^{[\text{stop}]} \equiv \int_{\nu}^{\infty} \{g(t)\}^2 w^{[\text{stop}]}(t) dt = \sum_{p,q} c_p c_q S_{p,q}$ と書ける. いま $w^{[\text{stop}]}(t) \equiv 1$ ならば, $S_{p,q} \equiv \int_{\nu}^{\infty} \frac{dx}{(t+a_p)(t+a_q)}$ の値は $p=q$ のときには $\frac{1}{\nu+a_p}$ となり, $p \neq q$ のときには $\frac{1}{a_p-a_q} \log \frac{\nu+a_p}{\nu+a_q}$ となる. 最小化条件を求めるために J の偏微分を 0 とおくと, $\frac{\partial J}{\partial c_p} = \sum_{q=1}^n S_{p,q} c_q + \frac{\lambda^{(L)}}{a_p} = 0$, $p=1, 2, \dots, n$, $\frac{\partial J}{\partial \lambda^{(L)}} = 1 - \sum_{p=1}^n \frac{c_p}{a_p} = 0$ となる. そこで, いま第 p 要素が $\frac{1}{a_p}$ であるベクトルを \mathbf{b} として, 連立 1 次方程式 $S\mathbf{c} = \mathbf{b}$ を解いて得られたベクトル \mathbf{c} を, $\sum_{p=1}^n \frac{c_p}{a_p}$ が 1 となるように定数を乗じて再規格化すれば求める解 \mathbf{c} が得られる. 連立 1 次方程式を解く際に係数の n 次対称行列 S の条件数は (n が増すと) 大きくなりがちなので, Rutishauser の Jacobi 法を用いて S の固有値分解を行い, 微小な固有値に対応する自由度を除去する切断操作を用いて, レゾルベントの係数を並べたベクトル \mathbf{c} のノルムの増大を抑制する.

4.1.2 区間 $[\nu, \infty)$ と区間 $[0, 1]$ での最小 2 乗法による構成

区間 $[\nu, \infty)$ と $[0, 1]$ はそれぞれフィルタの阻止域と通過域 (の近似) になることを意図している.

いま ν を 1 よりも大きい数として, $g(t)$ の 2 乗に非負の重み関数 $w^{[\text{stop}]}(t)$ を掛けて区間 $[\nu, \infty)$ に於いて積分した値を $J^{[\text{stop}]}$ とし, $1-g(t)$ の 2 乗に非負の重み関数 $w^{[\text{pass}]}(t)$ を掛けて通過域 $[0, 1]$ に於いて積分した値を $J^{[\text{pass}]}$ とする (重

み関数 $w^{[\text{pass}]}(t)$ は $[0, 1]$ での積分値が 1 になるようあらかじめ規格化されているとする). 微小な正の定数 ρ を導入して, Lagrangian を $J = J^{[\text{stop}]} + \rho J^{[\text{pass}]}$ とする. ここで, $J^{[\text{stop}]} \equiv \int_{\nu}^{\infty} \{g(t)\}^2 w^{[\text{stop}]}(t) dt = \sum_{p,q} c_p c_q S_{p,q}^{[\text{stop}]}$, および $J^{[\text{pass}]} \equiv \int_0^1 \{1-g(t)\}^2 w^{[\text{pass}]}(t) dt = \sum_{p,q} c_p c_q S_{p,q}^{[\text{pass}]} - 2 \sum_p b_p c_p + 1$ と書ける. パラメータ ρ を増やすと通過域に於ける近似が改善されるが, 阻止域に於ける近似が悪化する (ρ を減らすとその逆になる) というトレードオフの関係にあるため, ρ の値は通過域と阻止域での近似の程度のバランスをみて調節する.

いま $S = S^{[\text{stop}]} + \rho S^{[\text{pass}]}$ とおくと, 最小化条件から連立 1 次方程式 $S\mathbf{c} = \mathbf{b}$ が得られ, これを解いて係数 c_p を求める. 前と同様に, n 次対称行列 S は (n が増すと) 条件数が大きくなりがちなので, 連立 1 次方程式を解く際には固有値分解を用いて S の微小な固有値に対応する空間の自由度を除く切断操作を導入して係数 c_p の増大を抑制する.

4.2 実数シフトを用いたフィルタの作成例

4.2.1 作成したフィルタ (その 1)

伝達関数を $g(t) = \sum_{p=1}^n \frac{c_p}{t+a_p}$ とする. そうして $n=16$ として, シフト量は $-a_p = -10 \times 1.3^{-(p-1)}$, $p=1, 2, \dots, n$ と設定した (図 7). いま $\nu=3$ として, 区間 $[\nu, \infty)$ での $g(t)$ の 2 乗の積分の値 $J = \int_{\nu}^{\infty} \{g(t)\}^2 dt$ を拘束条件 $g(0)=1$ の下で最小化する係数 c_p を求めた (表 1). その際に n 次対称行列 S の階数は 15 であった. 係数 c_p の組を求める途中の過程では計算に四倍精度演算を用いたが, 得られた c_p は通常の倍精度に値を丸めて用いている. 求められた係数 c_p の絶対値の最大値は 1.49×10^2 であった. 得られたフィルタの特性に対して, $\mu=3$ とすると, $g_{\text{pass}}=3.468 \times 10^{-6}$, $g_{\text{stop}}=7.249 \times 10^{-15}$ となった. 伝達関数の大きさ $|g(t)|$ の対数値のグラフを図 8 に示す.

4.2.2 作成したフィルタ (その 2)

伝達関数は $g(t) = \sum_{p=1}^n \frac{c_p}{t+a_p}$ で, $n=16$ として, シフト量を $-a_p = -6 \times 1.3^{-(p-1)}$, $p=1, 2, \dots, n$ と設定した (図 9). いま $\nu=3$ として, 最小化する目的関数を $J = \int_{\nu}^{\infty} (g(t))^2 w^{[\text{stop}]}(t) dt + \rho \int_0^1 (1-g(t))^2 w^{[\text{pass}]}(t) dt$ とした. ただし積分の重み関数は, 区間 $[\nu, \infty)$ では $w^{[\text{stop}]}(t) \equiv 1$, 区間 $[0, 1]$ では $w^{[\text{pass}]}(t) \equiv \frac{1}{\pi} \frac{1}{\sqrt{t(1-t)}}$ とし, 定数 $\rho=10^{-20}$ の値を得られるフィルタの特性が良好となるように試行して決めた. 最小 2 乗法で c_p の値を決める際に, 連立 1 次方程式の係数である 16 次の対称行列 S は固有値の閾値 10^{-28} では階数が 13 となった (この閾値は S の有効な階数を定めるためのもので, 係数 c_p の絶対値の最大値が大き過ぎるときには, この閾値を上げて階数を減らすことで適度な大きさにする). 係数 c_p の組を求める過程では計算に四倍精度演算を用いたが, 得られた c_p の値は使用時には倍精度に丸めている. 得られた c_p の値は区間 $[0, 1]$ に

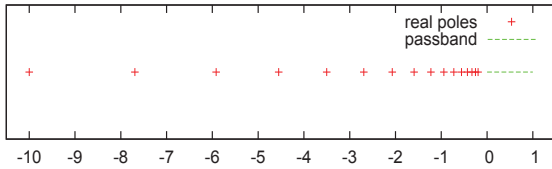


図 7 伝達関数の極の分布の例：今回作成した極が実数のみのフィルタ (その 1) ($n=16, \mu=3$)

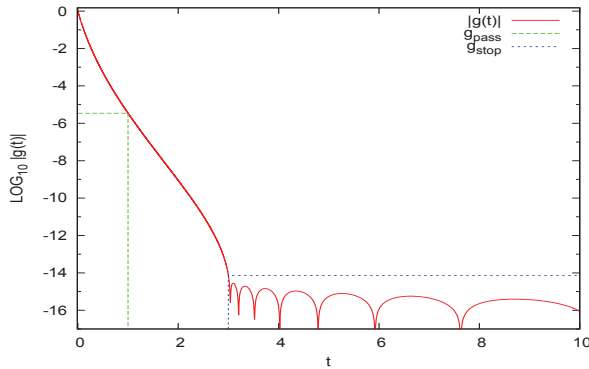


図 8 フィルタ (その 1) の伝達関数 $|g(t)|$ の対数値 ($n=16, \mu=3, g_{\text{pass}}=3.468 \times 10^{-6}, g_{\text{stop}}=7.249 \times 10^{-15}$)

表 1 フィルタ (その 1) の正規化座標 t でのシフト量 $-a_p$ と結合係数 c_p

p	シフト量 $-a_p$	結合係数 c_p
1	-1.0000 00000 00000 00 E+1	9.6048 64046 31508 22 E-6
2	-7.6923 07974 45986 73 E+0	-2.4175 79284 15423 10 E-4
3	-5.9171 60197 39388 66 E+0	3.1043 98643 41823 71 E-3
4	-4.5516 61857 25695 16 E+0	-2.6790 27412 37078 25 E-2
5	-3.5012 78480 16224 59 E+0	1.7175 41373 53764 66 E-1
6	-2.6932 91237 37567 68 E+0	-8.5171 83672 31578 34 E-1
7	-2.0717 62566 28078 02 E+0	3.2932 45960 77104 43 E+0
8	-1.5936 63570 97890 85 E+0	-9.7500 39877 98620 21 E+0
9	-1.2258 95099 56472 47 E+0	2.0931 03680 66355 57 E+1
10	-9.4299 62650 23300 45 E-1	-2.7880 77058 52754 49 E+1
11	-7.2538 17689 32460 45 E-1	6.5005 73350 26142 45 E+0
12	-5.5798 59965 68697 03 E-1	6.0304 69031 76852 71 E+1
13	-4.2922 00131 04232 43 E-1	-1.3707 28904 94552 41 E+2
14	-3.3016 92529 59945 59 E-1	1.4938 29318 97951 72 E+2
15	-2.5397 63577 46524 66 E-1	-8.5419 84876 74946 28 E+1
16	-1.9536 64362 01786 37 E-1	2.0414 95365 04261 31 E+1

於ける $g(t)$ の最大値が 1 となるように「再規格化」して用いた (表 2)。係数 c_p の絶対値の最大値は 1.09×10^2 である。得られたフィルタの特性に対して $\mu=3$ とすると、 $g_{\text{pass}}=1.047 \times 10^{-4}, g_{\text{stop}}=8.751 \times 10^{-12}$ となった。伝達関数の大きさ $|g(t)|$ の対数値のグラフを図 10 に示す。

4.3 いくつかの考察

シフトに実数を用いる場合には虚数を用いる場合とは異なりレゾルベントの作用を与えるための連立 1 次方程式は係数行列が実対称になる。さらにシフト量として最小固有値よりも小さい値を用いるなら係数行列は正定値となるので、疎行列の連立 1 次方程式を解く算法や解法ライブラリを選択する上では有利になる。しかしその反面、実数であ

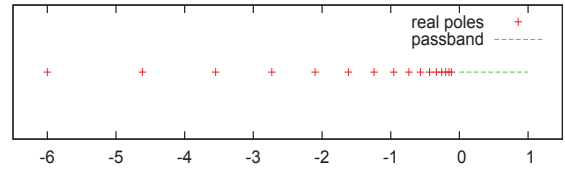


図 9 伝達関数の極の分布の例：今回作成した極が実数のみのフィルタ (その 2) ($n=16, \mu=3$)

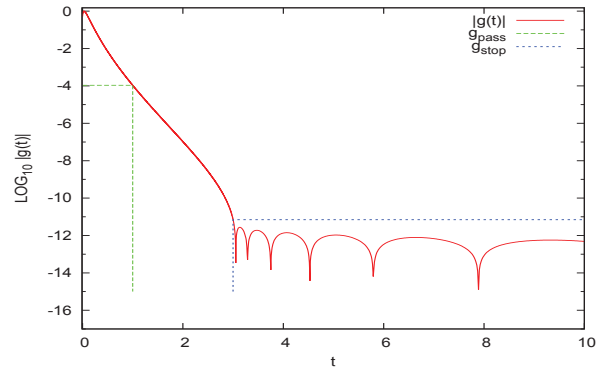


図 10 フィルタ (その 2) の伝達関数 $|g(t)|$ の対数値 ($n=16, \mu=3, g_{\text{pass}}=1.047 \times 10^{-4}, g_{\text{stop}}=8.751 \times 10^{-12}$)

表 2 フィルタ (その 2) の正規化座標 t でのシフト量 $-a_p$ と結合係数 c_p

p	シフト量 $-a_p$	結合係数 c_p
1	-6.0000 00000 00000 00 E+0	-4.1905 40378 30118 85 E-3
2	-4.6153 84784 67592 04 E+0	8.1348 52031 37557 93 E-2
3	-3.5502 96118 43633 20 E+0	-7.5871 04833 65196 20 E-1
4	-2.7309 97114 35417 10 E+0	4.4149 49097 85178 89 E+0
5	-2.1007 67088 09734 75 E+0	-1.7376 67541 02083 28 E+1
6	-1.6159 74742 42540 61 E+0	4.6683 24084 64327 88 E+1
7	-1.2430 57539 76846 81 E+0	-8.0858 18559 58570 66 E+1
8	-9.5619 81425 87345 11 E-1	7.1707 51443 70850 24 E+1
9	-7.3553 70597 38834 80 E-1	1.3557 04800 08579 11 E+1
10	-5.6579 77590 13980 27 E-1	-8.6406 19669 23325 43 E+1
11	-4.3522 90613 59476 27 E-1	1.7880 31490 50100 46 E+1
12	-3.3479 15979 41218 22 E-1	9.3209 79991 65761 31 E+1
13	-2.5753 20078 62539 46 E-1	-3.3069 17767 11109 85 E+1
14	-1.9810 15517 75967 35 E-1	-1.0721 89029 59509 82 E+2
15	-1.5238 58146 47914 79 E-1	1.0863 06746 36156 53 E+2
16	-1.1721 98617 21071 82 E-1	-3.0472 85100 75329 88 E+1

るシフト量を遷移域付近に配置できないので、(精度が限られた演算による情報落ちを防ぐためにレゾルベントの係数の大きさは制約せざるをえないこともあり) 伝達関数 $f(\lambda)$ の値は遷移域に於いて急峻な変化ができないのでフィルタの遮断特性はゆるやかになる。

フィルタの特性値 g_{pass} が 1 に比べて小さ過ぎる場合は、通過域に於ける伝達率の一様性が悪いので、縮小された固有値問題を解いて不変部分空間の近似基底を作る際に、丸め誤差の拡大で生じた異常な解が混ざることがある。そのような場合には、近似基底の B -正規直交性の低下をなるべく防ぐように慎重な方法で計算する (これらの困難は計算に用いる浮動小数点数の数値と演算の有効精度を増やせるならば原理的には回避できる)。また、 g_{stop} の値は g_{pass}

に比べて十分小さくしないと、不要な固有値を持つ固有ベクトルが混入する割合が増えてノイズが多くなり、固有対が十分な精度で求まらなくなる。また遷移域の幅 $\mu - 1$ が広いと遷移域に固有値が入る固有対の個数もそれだけ多くなりがちであるが、フィルタで濾過するベクトルの個数は通過域または遷移域に固有値がある固有対の個数よりも多くする必要があるので、計算量と記憶量の面からはなるべく μ を 1 に近くできることが望ましいが、今の場合にはそれは難しく今回の 2 通りのフィルタはどちらも $\mu = 3$ である。

5. 作成したフィルタを用いた対角化の実験例

シフト量が実数だけのレゾルベントの線形結合であるフィルタを作成し、それを用いたフィルタ対角化法で近似固有対がうまく構成できるかについて数値実験を行って調べた。その例を 2 つほど紹介する。

まず、以下の実験で用いた計算機システムの諸元をここで述べておく。CPU は Intel Core i7-5960X (3.0GHz, 8 コア, 20MB の共有 L3 キャッシュで, Turbo モードと Hyperthread 機能はオフにしてある) で, 主記憶は Quad Channel 動作で 8G バイトのモジュール (DDR4-17000) が 8 本で合計容量 64G バイト。OS は CentOS 7.0 (64 ビット) である。コンパイラには Intel Fortran ver.15.0.0 を用いた。コンパイラへのオプションは最適化には `-fast` を, OpenMP を用いたスレッド並列化には `-openmp` を与えた。この対角化の実験の計算に用いた浮動小数点数とその演算は全て IEEE-754 の 64 ビット倍精度である (この Haswell アーキテクチャーの CPU は Intel AVX2 拡張命令セットの融合乗加算命令を持つ)。

5.1 例題 1 (フィルタ (その 1) を使用)

以下の例題で扱う実対称定値一般固有値問題 $A\mathbf{v} = \lambda B\mathbf{v}$ は, 立方体領域 $[0, \pi] \times [0, \pi] \times [0, \pi]$ に於ける 3 次元 Laplace 偏微分作用素の零 Dirichlet 境界条件の固有値問題 $-(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2})\Psi(x, y, z) = \lambda \Psi(x, y, z)$ を有限要素法 (FEM) で離散化して得られたものである。但し, 有限要素内の基底関数には各座標軸方向の区分線形関数の直積を用いた。立方体領域を各座標軸方向に 26 等分したものを有限要素として用いたので, 行列の次数は $N = 25 \times 25 \times 25 = 15,625$ となる。一般固有値問題の係数行列 A, B は帯対称で, FEM の基底関数の適切な番号付けにより (対角線を含めない) 下帯幅は $1 + 25 + 25^2 = 651$ になる。

Laplace 偏微分作用素の零 Dirichlet 境界条件の固有値問題を FEM により離散化して得られる行列の一般固有値問題の固有値は必ず正の実数なので, 固有値分布の下端付近の区間 $[0, 30]$ に固有値がある近似対 (真の個数は 54 個) を求めてみた。ちなみに, この離散化で得られる行列の一般固有値問題の固有値の厳密な値は, 簡単な数式の計算によ

り求められる。

今の場合にはシフトが実数であるレゾルベントの作用を実現するための連立 1 次方程式の係数行列は実対称帯で正定値であるので, 枢軸選択を行わない帯用の修正 Cholesky 分解 (LDL^T 分解) を用いて数値的に安定に解くことができる。他方で, 逆反復法を用いて固有対を改良する際に解くべき連立 1 次方程式の係数 $A - \lambda B$ は実対称帯行列であるが, 一般には (近似固有値である実数のシフト量 λ が最小固有値より小さくしなければ) 正定値ではないので, 数値安定性を重視して帯行列用の部分枢軸選択を行う LU 分解を用いて対称性を利用せずに解いている。そのため LU 分解では下帯幅は 651, 上帯幅はその 2 倍の 1,302 の記憶を確保して計算を行った (帯行列の連立 1 次方程式の解法には直接法である修正 Cholesky 分解や LU 分解を用いたが, 今回の FEM の離散化で得られる帯行列は帯の内部が著しく疎であるが, それをあたかも密であるかのように扱って分解していることを注意しておく)。

この例題 1 で用いたフィルタ (その 1) は実数シフトのレゾルベントの線形結合で, その構成法は既に上で記述したものである。シフトの個数は $n=16$, フィルタの特性値は $\mu=3$, $g_{\text{pass}}=3.468 \times 10^{-6}$, $g_{\text{stop}}=7.249 \times 10^{-15}$ である。乱数から生成した $m=300$ 個のベクトルを B -正規直交化して構成したベクトルの組 X をフィルタへの入力組とした。論文 [2],[3],[7] に記述された方法にほぼ従って (但し異なる点もあるので詳細を付録に記述), フィルタの出力の組 Y から不変部分空間の近似基底 Z をフィルタの伝達特性を考慮して構成し, それから Rayleigh-Ritz 法で固有対を取り出すフィルタ対角化法による実験の例を示す。

いま m 次対称行列 $\beta = X^T B Y$ の固有値をマシン精度シロンの 100 倍の値 2.2×10^{-14} を閾値にして切断すると β の有効階数は 259 となったので, 既に入力ベクトルの個数 $m=300$ でほぼ飽和に達しているとみなせる。図 11 のグラフは, β の固有値を減少順に横軸に並べて固有値の絶対値を対数で縦軸にプロットしたものである (固有値の符号が正の点は赤色で, 負の点は青色でプロットして表し, 緑色の水平線で切断に用いた閾値 2.2×10^{-14} の位置を表している)。

図 12 は, フィルタ作用素 \mathcal{F} の近似固有値 ϕ の分布を示すグラフで, 横軸に ϕ の値に減少順に振った番号をとり, 縦軸に ϕ の絶対値の対数をとってプロットしたものである (ϕ の符号が正の点は赤色で, 負の点は青色で描かれている)。第 54 番目までが通過域に対応している。 ϕ が $g_{\text{pass}}=3.468 \times 10^{-6}$ 以上のものだけを集めると, 真の固有対の個数と一致する 54 個の基底が得られた (基底を 54 個にとった場合の近似対の固有値はすべて $[0, 30]$ にあった)。閾値を半分以下に下げ 1.734×10^{-6} にすると (グラフの緑色の水平線はこの閾値を表す) 基底は 60 個になり, 6 重に縮退した固有値 30.79 を持つ余分な近似対も含めて求まった。

図 13 は、フィルタ対角化法で得られた近似対のうちで固有値が $[0, 30]$ にあるものをグラフの折れ線 IT0 で、横軸に固有値をとり、縦軸に残差のノルム Δ の値をとってプロットしたものである (近似対 (λ, \mathbf{v}) のベクトルは、 B -正規直交条件 $\mathbf{v}^T B \mathbf{v} = 1$ で規格化されている。そのとき、固有対の残差を $\mathbf{r} = A\mathbf{v} - \lambda B\mathbf{v}$ と定義し、残差のノルムを $\Delta = \sqrt{\mathbf{r}^T B^{-1} \mathbf{r}}$ により定義する)。このグラフからフィルタ対角化法により得られた近似対の残差のノルム Δ は 10^{-9} から 10^{-4} 程度で、そのことから近似固有値の相対精度は少なくとも 4 桁から 5 桁程度はあることがわかる (後で示すように実際の精度はもっと高い)。同じグラフ中の折れ線 IT1, IT2 はそれぞれ、フィルタ対角化法で得られた近似対に対して Rayleigh 商逆反復によりそれぞれ 1 回, 2 回改良を加えて得られた近似対について残差のノルムの対数をプロットしたものである。

この行列の一般固有値問題は簡単な数式の計算で固有値の厳密な値を求めることができる。そこで (Rayleigh 商逆反復法による改良前の) フィルタ対角化法で得られた区間 $[0, 30]$ に固有値がある近似対について、横軸に近似固有値の値、縦軸にその真値からの絶対誤差の大きさの対数値をとってグラフにプロットした (図 14)。得られた近似固有値の絶対誤差の最大値は 6.36×10^{-11} 、相対誤差の最大値は 2.16×10^{-12} であった。

現在の実装では OpenMP を用いて 8 スレッドで CPU 内の 8 個のコアを用いて並列に計算を実行した場合は、フィルタ対角化全体の経過時間は 10.09 秒であり、(不要な対も含む) 近似対 60 個のすべてについて逆反復を 2 回適用する経過時間全体は 508.3 秒であった。また 1 スレッドで計算を実行した場合のフィルタ対角化の経過時間は 58.16 秒であった (これは 8 スレッドの場合の時間の 5.76 倍である) (表 3)。たとえば「フィルタの適用」の処理ステップでは、用いたフィルタを構成する $n = 16$ 個のレゾルベントの作用を適用するには全部で 8 個のスレッドが各スレッドごとに 2 個ずつを処理している。

今回のフィルタの適用には、シフト行列 $A - \tau_p B$ の実対称性と正定値性を利用して対称帯行列用の枢軸選択をしない修正 Cholesky 分解を合計で $n = 16$ 通り実行している。その後の 60 個の近似対を 2 回ずつ改良逆反復法で修正する場合は、各近似固有値によるシフト行列は実対称であるが正定値ではないので、枢軸選択を行い対称性を利用しない帯行列用の LU 分解を合計で $60 \times 2 = 120$ 通り実行している。この例題 1 の 8 スレッド計算では、「フィルタの適用」ステップはレゾルベントの作用を与える帯行列の連立 1 次方程式を係数が対称正定値の場合には数値安定性が保証される「ピボット交換を行わない対称帯用の修正 Cholesky 分解」を用いて解いており、経過時間は 7.36 秒であった。比較のために、「フィルタの適用」ステップでレゾルベントの作用を与える帯行列の連立 1 次方程式を、係

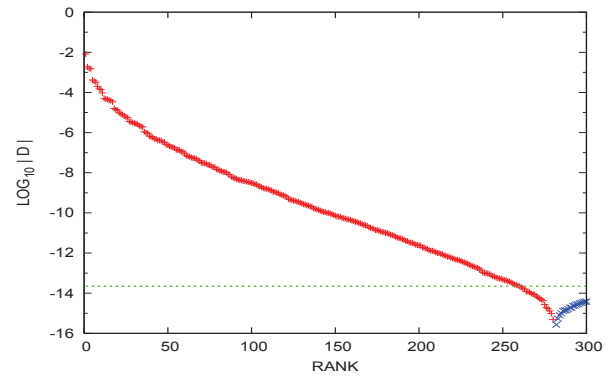


図 11 例題 1: β の固有値の分布 ($m=300$, 閾値 2.2×10^{-14})

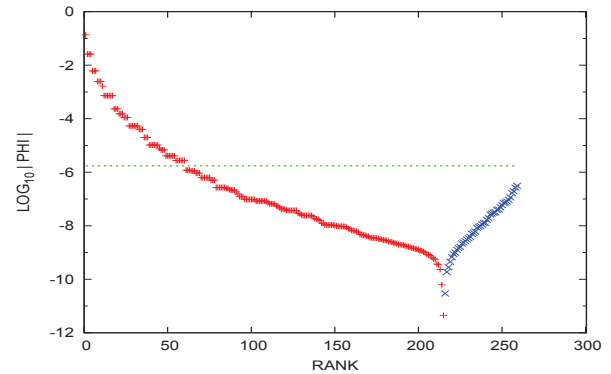


図 12 例題 1: $|\phi|$ の値の分布 ($m=300$, 閾値 1.734×10^{-6})

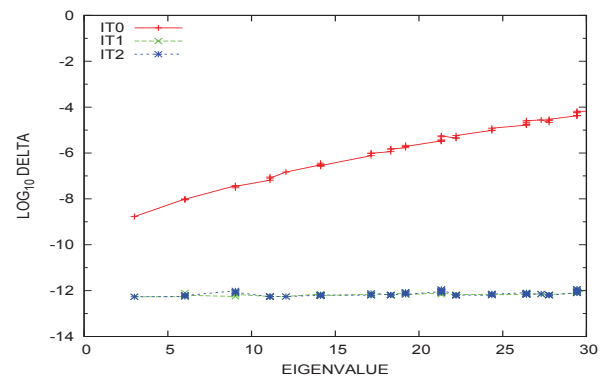


図 13 例題 1: 近似対の残差のノルム Δ (区間 $[0, 30]$, $m=300$)

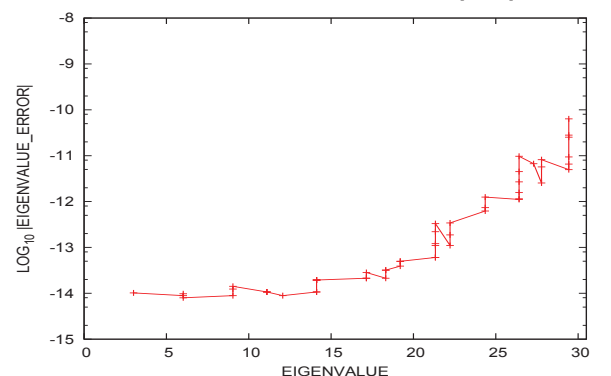


図 14 例題 1: 近似固有値の絶対誤差 ($m=300$, ϕ の閾値 1.734×10^{-6})

数が対称正定値ではない場合にも数値安定である「片側ピボット交換付きの非対称な帯行列用の LU 分解」を用いて解くと経過時間は 80.43 秒であった。シフトを最小固有値

表 3 例題 1：経過時間（秒）の内訳

	8 スレッド	1 スレッド
フィルタ対角化全体	10.09	58.16
- 乱数ベクトル生成	0.02	0.02
- 正規直交化	0.71	5.03
- フィルタの適用	7.36	43.05
- 不変部分空間の基底作成	1.73	8.26
- Rayleigh-Ritz 法	0.26	1.80
Rayleigh 商逆反復 2 回分	508.35	669.25

よりも小さい実数に選んだことで、レゾルベントの作用を与える連立 1 次方程式の係数が対称正定値となり、それにより「フィルタの適用」ステップは経過時間が短縮されて、この例では約 10.9 倍高速に計算ができたことになる。

5.2 例題 2（フィルタ（その 2）を使用）

この例題 2 も、例題 1 と同じく立方体領域 $[0, \pi] \times [0, \pi] \times [0, \pi]$ に於ける 3 次元 Laplace 偏微分作用素の零 Dirichlet 境界条件での固有値問題を FEM で離散化して得られる実対称定値一般固有値問題に対して、固有値が下端側にある固有対を求める例である。有限要素内の基底関数として各座標軸方向の区分線形関数の直積を用いることも例題 1 と同じであるが、今度は立方体領域の各座標軸方向を 41 等分した有限要素を用いた。離散化で得られる一般固有値問題の係数行列 A, B の次数は $N = 40 \times 40 \times 40 = 64,000$ である。また FEM の基底関数に適切に番号を付けると（対角線を含めない）下帯幅は $1 + 40 + 40^2 = 1,641$ になる。そうして、固有値が固有値分布の下端側の区間 $[0, 30]$ にある近似固有対（真の個数は 54 個）を求めてみた。この固有値問題の固有値の真値も簡単な数式の計算で求められる。

用いたフィルタ（その 2）はシフトが実数であるレゾルベントの線形結合で、その構成法は既に上で記述したものである。シフトの個数は $n=16$ 、フィルタの特性値は $\mu=3$ とすると、 $g_{\text{pass}}=1.047 \times 10^{-4}$ 、 $g_{\text{stop}}=8.751 \times 10^{-12}$ である。乱数を用いて作った $m=300$ 個のベクトルを B -正規直交化したものを入力ベクトルの組 X として作り、 X にフィルタを作用させて出力ベクトルの組 Y を作る。それら X と Y から必要な固有値に対応する不変部分空間を近似する基底 Z を構成して、Rayleigh-Ritz 法に与えるフィルタ対角化法の実験例を示す。

m 次対称行列 $\beta = X^T B Y$ の固有値をマシンイプシロンの 100 倍の値 2.2×10^{-14} を閾値にして切断すると、 β の有効階数は 299 であった。つまり入力ベクトルの個数 $m=300$ でちょうど飽和に達した程度である。図 15 のグラフは、この β の固有値を減少順に横軸に並べて固有値の絶対値を対数で縦軸にプロットしたものである（この問題の場合には β の固有値はすべて正で、赤色の点で示してある。緑色の水平線で切断に用いた閾値 2.2×10^{-14} の位置を表している）。

図 16 のグラフには、フィルタ作用素 \mathcal{F} の近似固有値 ϕ の分布を ϕ の値の減少順に番号を振ったものを横軸にとり、 ϕ の絶対値の対数を縦軸にプロットしたものを示す（ ϕ の符号が正の点は赤色で、負の点は青色で描いてある）。元の一般固有値問題の固有値を λ とすると、 ϕ は $f(\lambda)$ の近似値であり、第 54 番目までが通過域に対応する。そこで ϕ が $g_{\text{pass}}=1.047 \times 10^{-4}$ 以上のものだけを集めると、真の固有対の個数と一致する 54 個の基底が得られた（この 54 個の基底から得られる近似対の固有値はすべて $[0, 30]$ にあった）。安全のため余裕を設けて閾値を半分の値 5.23×10^{-5} に下げると（グラフの緑色の水平線はこの閾値を表す）基底は 60 個となり、6 重に縮退した固有値 30.32 を持つ余分な近似対も含めて求まる。

図 17 のグラフは、フィルタ対角化法により得られた（Rayleigh 商逆反復法による改良前の）近似固有対のうちで固有値が $[0, 30]$ にあるものを折れ線 IT0 で、横軸に近似固有値を、縦軸に残差のノルム Δ の値の対数をとってプロットしたものである。このグラフからフィルタ対角化法による近似対の残差のノルムは 10^{-7} から 10^{-4} の程度であること、そのことから近似固有値の相対精度は少なくとも 5 桁程度であることがわかる（実際の精度はもっと高い）。同じグラフ中の折れ線 IT1, IT2 はそれぞれ、フィルタ対角化法で得られた近似対に対して Rayleigh 商逆反復によりそれぞれ 1 回、2 回改良を加えて得られた近似対について残差のノルムの対数をプロットしたものである。

固有値の真値との比較を行い、（Rayleigh 商逆反復法による改良前の）フィルタ対角化法で得られた近似固有対で固有値が $[0, 30]$ にあるものについて、横軸に近似固有値の値、縦軸にその絶対誤差の大きさの対数値をとってプロットした（図 18）。得られた近似固有値の絶対誤差の最大値は 2.36×10^{-9} で、相対誤差の最大値は 8.08×10^{-11} であった。

表 4 例題 2：経過時間（秒）の内訳

	8 スレッド	1 スレッド
フィルタ対角化全体	380.33	768.85
- 乱数ベクトル生成	0.08	0.08
- 正規直交化	15.79	51.66
- フィルタの適用	332.53	622.19
- 不変部分空間の基底作成	25.67	76.53
- Rayleigh-Ritz 法	6.27	18.38
Rayleigh 商逆反復 2 回分	13,381.83	36,227.56

OpenMP を用いて 8 スレッドの並列計算を CPU の 8 コアで行った場合の経過時間は、フィルタ対角化全体が 380.3 秒で、フィルタ対角化で得られた（余分な 6 個の対を含む）近似対 60 個の全てに対して Rayleigh 商付き逆反復による改良を 2 回ずつ行うのに掛かった時間全体は 13,382 秒であった。同じことを 1 スレッドで行った場合の経過時間は、

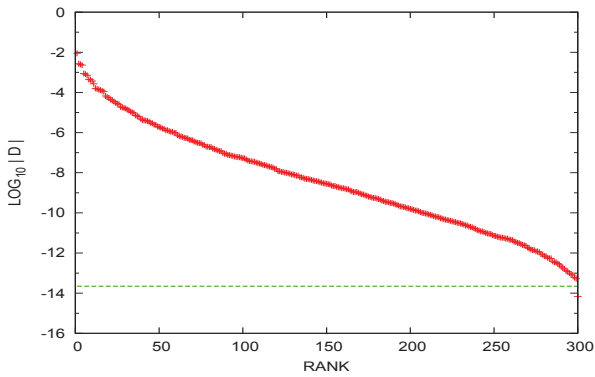


図 15 例題 2: β の固有値の分布 ($m=300$, 閾値 2.2×10^{-14})

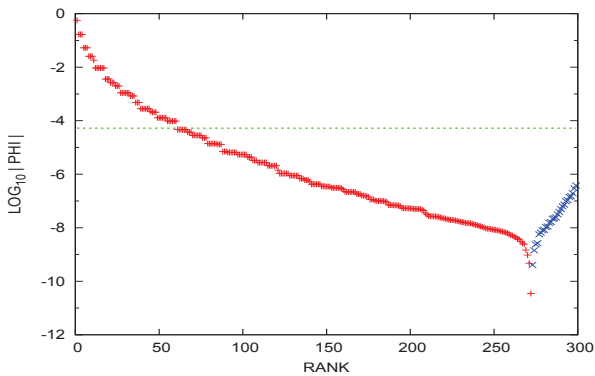


図 16 例題 2: $|\phi|$ の値の分布 ($m=300$, 閾値 5.23×10^{-5})

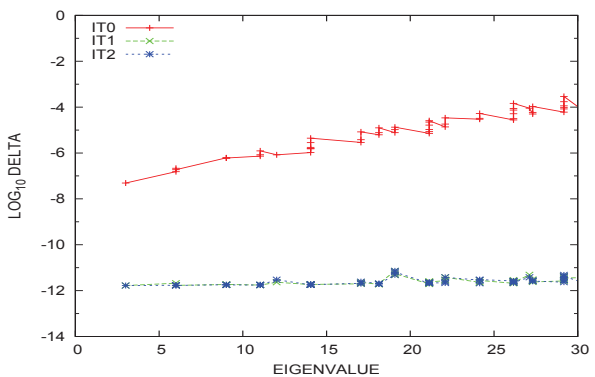


図 17 例題 2: 近似対の残差のノルム Δ (区間 $[0, 30]$, $m=300$)

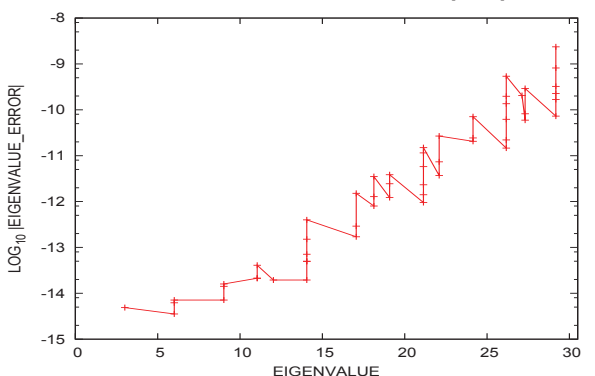


図 18 例題 2: 近似固有値の絶対誤差 ($m=300$, ϕ の閾値 5.23×10^{-5})

フィルタ対角化が 768.9 秒で、フィルタ対角化で得られた近似対 60 個の全てに対して Rayleigh 商付き逆反復による改良を 2 回ずつ行うのに掛かった時間全体は 36,228 秒で

あった (表 4)。

この例題 2 の 8 スレッド計算では、「フィルタの適用」ステップは「ピボット交換を行わない対称帯用の修正 Cholesky 分解」を用いて解いており、経過時間は 332.53 秒であった。比較のために、「フィルタの適用」ステップを「片側ピボット交換付きの非対称な帯行列用の LU 分解」を用いて解くと経過時間は 1960.12 秒であった。連立 1 次方程式の係数が帯対称正定値であることを利用することで、この例では「フィルタの適用」ステップの計算を約 5.9 倍高速に行なえたことになる。

5.3 例題 3 (フィルタ (その 1) を使用)

この例題 3 も、立方体領域 $[0, \pi] \times [0, \pi] \times [0, \pi]$ に於ける 3 次元 Laplace 偏微分作用素の零 Dirichlet 境界条件での固有値問題を FEM で離散化して得られた実対称定値一般固有値問題に対して、固有値が下端側にある固有対を求める例である。有限要素内の基底関数として各座標軸方向の区分線形関数の直積を用いることも同じであるが、今度は立方体領域の各座標軸方向をそれぞれ 31, 41, 51 に等分した有限要素を用いた点が例題 1 や 2 とは異なっている。離散化で得られる一般固有値問題の係数行列 A, B の次数は $N = 30 \times 40 \times 50 = 60,000$ である。また FEM の基底関数に適切に番号を付けると (対角線を含めない) 下帯幅は $1 + 30 + 30 \times 40 = 1,231$ になる。そうして、固有値が固有値分布の下端側の区間 $[0, 30]$ にある近似固有対 (真の個数は 54 個) を求めてみた。この固有値問題の固有値の真値も簡単な数式の計算で求められる。この問題の場合には固有値には縮退が無い。

用いたフィルタ (その 1) はシフトが実数であるレゾルベントの線形結合で、その構成法は既に上で記述したものである。シフトの個数は $n=16$ 、フィルタの特性値は $\mu=3$ とすると、 $g_{\text{pass}}=3.468 \times 10^{-6}$ 、 $g_{\text{stop}}=7.249 \times 10^{-15}$ である。乱数を用いて作った $m=300$ 個のベクトルを B -正規直交化したものを入力ベクトルの組 X として作り、 X にフィルタを作用させて出力ベクトルの組 Y を作る。それら X と Y から必要な固有値に対応する不変部分空間を近似する基底 Z を構成して、Rayleigh-Ritz 法に与えるフィルタ対角化法の実験例を示す。

マシンイプシロンの 100 倍の値 2.2×10^{-14} を閾値にして m 次対称行列 $\beta = X^T B Y$ の固有値を切断すると、 β の有効階数は 252 であった。つまり入力ベクトルの個数 $m=300$ では既にほぼ飽和に達している。図 11 のグラフは、この β の固有値を降順に横軸に並べて固有値の絶対値を対数で縦軸にプロットしたものである (固有値の符号が正の点は赤色で、負の点は青色でプロットし、緑色の水平線で切断に用いた閾値 2.2×10^{-14} の位置を表している)。図 20 のグラフには、フィルタ作用素 \mathcal{F} の近似固有値 ϕ の分布を ϕ の値の減少順に番号を振り、その絶対値の対数でプロットし

たものを示す (ϕ の符号が正の点は赤色で、負の点は青色で描いてある). 元の一般固有値問題の固有値を λ とすると, ϕ は $f(\lambda)$ の近似値であるので, 第 54 番目までが通過域に対応する. そこで ϕ が $g_{\text{pass}}=3.468 \times 10^{-6}$ 以上のものだけを集めると, 真の固有対の個数と一致する 54 個の基底が得られた (この 54 個の基底から得られる近似対の固有値はすべて $[0, 30]$ にあった). 安全のため余裕を設けて閾値を半分の値 1.734×10^{-6} に下げると (グラフの緑色の水平線はこの閾値を表す) 基底は 60 個となり, 固有値 30.20694, 30.21245, 30.31319, 30.32130, 30.54491, 30.54751 を持つ余分な近似対も含めて求まる.

フィルタ対角化法により得られた (Rayleigh 商逆反復法による改良前の) 近似固有対のうちで固有値が $[0, 30]$ にあるものをグラフ (図 21) に折れ線 IT0 で, 横軸に近似固有値を, 縦軸に残差のノルム Δ の値の対数をとってプロットした. このグラフからフィルタ対角化法による近似対の残差のノルムは 10^{-8} から 10^{-4} の程度であること, そのことから近似固有値の相対精度は少なくとも 5 桁程度であることがわかる (実際の精度はもっと高い). 同じグラフ中の折れ線 IT1, IT2 はそれぞれ, フィルタ対角化法で得られた近似対に対して Rayleigh 商逆反復によりそれぞれ 1 回, 2 回改良を加えて得られた近似対について残差のノルムの対数をプロットしたものである.

固有値の真値との比較を行い, (Rayleigh 商逆反復法による改良前の) フィルタ対角化法で得られた近似固有対で固有値が $[0, 30]$ にあるものについて, 横軸に近似固有値の値, 縦軸にその絶対誤差の大きさの対数値をとってプロットした (図 22). 得られた近似固有値の絶対誤差の最大値は 3.99×10^{-10} で, 相対誤差の最大値は 1.31×10^{-11} であった.

表 5 例題 3: 経過時間 (秒) の内訳

	8 スレッド	1 スレッド
フィルタ対角化全体	199.86	487.82
- 乱数ベクトル生成	0.07	0.07
- 正規直交化	6.65	35.19
- フィルタの適用	179.90	385.00
- 不変部分空間の基底作成	10.72	54.62
- Rayleigh-Ritz 法	2.53	12.95
Rayleigh 商逆反復 2 回分	7,120.01	17,382.70

OpenMP を用いて 8 コアの CPU によるスレッド並列計算の経過時間は, 8 スレッドで行った場合はフィルタ対角化全体では 199.9 秒で, フィルタ対角化で得られた (余分な 6 個の対を含む) 近似対 60 個の全てに対して Rayleigh 商付き逆反復による改良を 2 回ずつ行うのに掛かった時間全体が 7,120 秒であった. 同じ計算を 1 スレッドで行った場合は, フィルタ対角化が 487.8 秒で, フィルタ対角化で得られた近似対 60 個の全てに対して Rayleigh 商付き逆反復

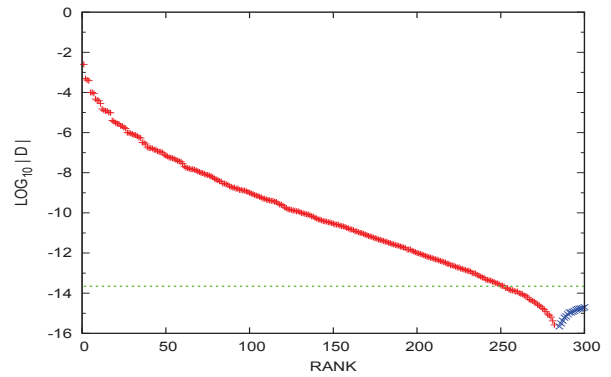


図 19 例題 3: β の固有値の分布 ($m=300$, 閾値 2.2×10^{-14})

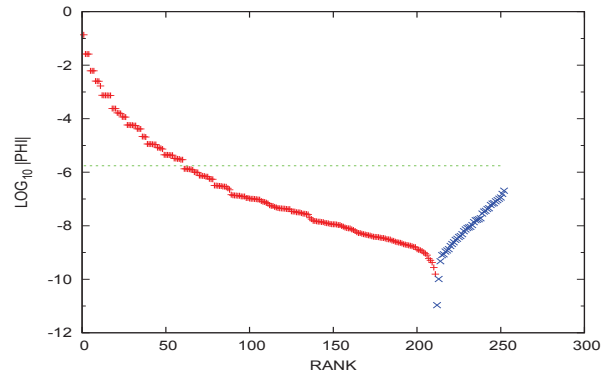


図 20 例題 3: $|\phi|$ の値の分布 ($m=300$, 閾値 1.73×10^{-6})

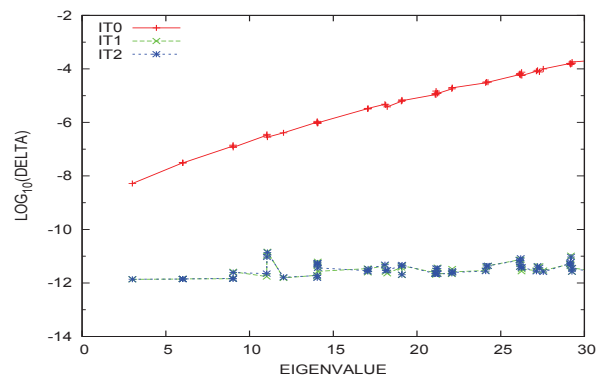


図 21 例題 3: 近似対の残差のノルム Δ (区間 $[0, 30]$, $m=300$)

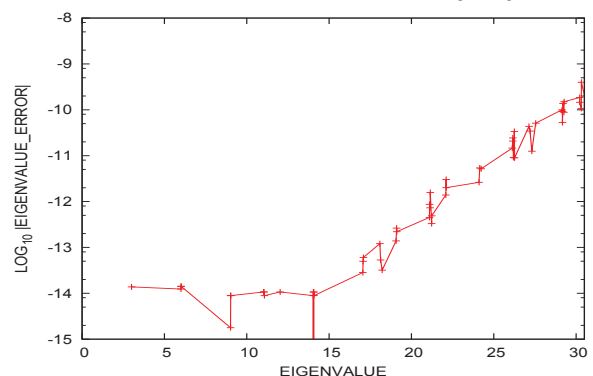


図 22 例題 3: 近似固有値の絶対誤差 ($m=300$, ϕ の閾値 1.73×10^{-6})

による改良を 2 回ずつ行うのに掛かった時間全体は 17,383 秒であった (表 5).

ベクトル $m = 300$ 個への $n = 16$ の「フィルタの適用」

を行なうステップの経過時間の値は、385.0 秒 (1 スレッド), 197.3 秒 (2 スレッド), 168.4 秒 (4 スレッド), 179.9 秒 (8 スレッド) となっていて、2 スレッド程度で性能がほぼ飽和し、4 スレッドから 8 スレッドにするとかえって経過時間が増加している。また、近似対 60 個への Rayleigh 商逆反復を各対に対して 2 回ずつ行なうステップの経過時間では、17,383 秒 (1 スレッド), 9,929 秒 (2 スレッド), 6,920 秒 (4 スレッド), 7,120 秒 (8 スレッド) となっていて、1 スレッドから 2 スレッドへの性能の向上比は 1.75 倍である程度は良いが、2 スレッドから 4 スレッドの性能の向上比は 1.43 倍と鈍化していて、4 スレッドから 8 スレッドにするとかえって経過時間が増加している。スレッド数の増加に伴う性能の向上が良くない理由としては、現在の実装ではフィルタの適用に現れる 16 個のレゾルベントの作用を与える連立 1 次方程式の係数は次数が $N = 60,000$ で半帯幅が 1,231 の帯行列であるが、それを帯行列用の極めて標準的な修正 Cholesky 法で解いていて、メモリ参照の局所性が高くないこと。また Rayleigh 商逆反復法についても、同じ次数と帯幅の帯行列でシフトが 60 通り異なるものについて、それぞれ対称性を利用せずに枢軸交換を行なって上帯幅を 2 倍に増やしながら標準的な (ブロック化などを行っていない) LU 分解を用いて解いていて、やはりメモリ参照の局所性が高くないこと。そのためマルチコアの CPU が 1 個で行列計算を並行に複数個をスレッド並列で行なうと、CPU とメモリ間のメモリバンド幅が不足するであろうことがあげられる。

6. まとめ

今回の研究では、与えられた実対称定値一般固有値問題 $A\mathbf{v} = \lambda B\mathbf{v}$ に対して、実数 a を最小固有値の下限とするときに、固有値分布の下端付近の区間 $[a, b]$ を通過域とするフィルタ作用素 \mathcal{F} を、 a より小さい実数をシフト量とするレゾルベントの線形結合で構成する方法を具体的に示した。ただし今回の方法は、まず先にレゾルベントのシフト量の組を与えておいてから、それらの線形結合の係数をフィルタの特性がなるべく良いものとなるように最小 2 乗法で決めるものである。レゾルベントの個数 n が与えられたときのシフト量の最適な配置もしくは系統的に良い配置の選択を見出すことは、今後の研究課題である。

得られるフィルタの伝達関数は、極を最小固有値より小さい実数の範囲に制限したことにより、極を複素数の範囲から選べる場合 (例えば Chebyshev フィルタや elliptic フィルタなど) と比べると、遷移域に於ける伝達率の変化をあまり急峻にはできないので遷移域の幅が広がってしまい、また通過域での伝達率の一様性も悪い。このためフィルタで濾過されたベクトルの組から精度の良い固有対の近似を得ることは困難になると予想される。しかし、そのような特性の良くないフィルタを用いる場合であっても、

「フィルタへの入力ベクトルの組と出力ベクトルの組を併せて分析し、フィルタの伝達特性に基づいて不変部分空間の近似基底を構成する方法」[2], [3], [7] を適用することで困難の程度を軽減できると予想した。そこで例題として、「有限要素法 (FEM) の離散化で生じた行列の対称定値一般固有値問題」の下端付近の固有値を持つ固有対を求める問題に対して、今回の構成法によるフィルタを用いてフィルタ対角化法を適用してみたところ、ある程度良い近似解を得ることができた。このことから、今回のフィルタ構成法が一応の妥当性を持つこと、およびフィルタ特性に基づいて不変部分空間の近似基底を構成する方法 [2], [3], [7] が有用であることを示せた。

固有値問題を実数のシフトだけを用いて解く試みについては、最近の文献 [8] にも記述がある。

付 録

Rayleigh-Ritz 法に与えるための不変部分空間の近似基底の構成法を記述した著者の最近の論文 [7] は、元の一般固有値問題の係数行列が複素エルミートの場合で、中間固有値を求めることを想定してフィルタの伝達関数の極 (レゾルベントのシフト量) は共役な複素対の組を用いる場合についてであるが、その方法を今回扱っている係数行列が実対称で伝達関数の極が (最小固有値よりも小さい) 実数である状況に沿って書き直したものを、以下に付録として記述する (同論文には行列 β の固有値を切断する閾値の設定法に対する考察も若干の実験例と共に記述してあるが、それについては省略する)。

A.1 Rayleigh-Ritz 法に与える基底の構成法

フィルタを作用させた後のベクトルの組 Y の張る空間は、「通過域 $[a, b]$ あるいは遷移域に固有値のある固有ベクトル」の張る空間を近似している。そこで、Rayleigh-Ritz 法 (R-R 法) を用いて Y の張る空間内で元の一般固有値問題 (GEVP) の近似固有対を求めようとするのであるが、フィルタの持つ伝達特性の性質から、固有値が阻止域にあるもしくは通過域から離れた遷移域にある固有ベクトルは強く減衰を受けるので、 Y の数値的な階数は通常は多く落ちている (しかもベクトルの個数 m を増せばそれだけより多くの階数が落ちる傾向がある)。そのような数値的な条件の悪い基底 Y に対して R-R 法を直接適用すると、得られる結果もまた数値的に不安定になる。

そこで一種の正則化として、 Y の張る空間の内からあらかじめ数値的な線形独立性の高い基底 Z を選び出す処理を行う。いま通過域 $[a, b]$ (のある近傍) に固有値を持つ不変部分空間を \mathcal{I} とする。そのとき、 Y の張る空間の中に制限された B -正規直交基底 Z を適切に構成して、 Z の張る空間が \mathcal{I} をなるべく良く近似するようにできる。

そのようにして構成した基底 Z に対して R-R 法を適用

して得られる Ritz 対を元の GEVP の近似固有対とする。具体的には、 Z は B -正規直交 ($Z^T B Z = I$) であるので、次数の小さい実対称行列 $\tilde{A} = Z^T A Z$ を作り、その両側直交変換による対角化 (固有値分解) を $\tilde{A} = U D U^T$ とすると、Ritz 対は ZU の列ベクトル (Ritz ベクトル) とそれに対応する D の対角要素 (Ritz 値) の対として得られる。この方法により得られる近似固有値 (Ritz 値) は実対称行列 \tilde{A} の固有値なので実数である。

A.1.1 不変部分空間 \mathcal{I} の近似基底 Z の構成

フィルタ作用素 \mathcal{F} がレゾルベントの線形結合の場合には、元の GEVP: $A\mathbf{v} = \lambda B\mathbf{v}$ の固有対を (λ, \mathbf{v}) とするとき、ベクトル \mathbf{v} は同時に \mathcal{F} の固有ベクトルであり $\mathcal{F}\mathbf{v} = f(\lambda)\mathbf{v}$ も満たす。ただし $f(\lambda)$ はフィルタの伝達関数である。その逆の命題「 \mathcal{F} の任意の固有ベクトルは元の GEVP の固有ベクトルである」は成り立たないが、「 \mathcal{F} の (任意の) 不変部分空間は元の GEVP の不変部分空間である」は成り立つ。

この性質を利用すると、「 $[a, b]$ のある近傍」に固有値を持つ元の GEVP の不変部分空間を \mathcal{I} とするとき、以下の性質を満たすベクトルの組 Z を構成できる：

- (1) 組 Z のベクトルは Y の張る空間内にある。
- (2) Z の張る空間は、不変部分空間 \mathcal{I} のよい近似である。
- (3) 組 Z は B -正規直交系である。

そのように構成された組 Z に対して R-R 法を適用すれば、固有値が「 $[a, b]$ のある近傍」にある元の GEVP の近似固有対を一斉にすべて求めることができる。

組 Z の構成は次のようにする。まず次の副節 (A.1.2) の方法を用いることで、ベクトル \mathbf{v} を Y の張る空間内に制限して得られる作用素 \mathcal{F} の近似固有対 (ϕ, \mathbf{v}) のなかから、固有値 ϕ の値が区間 $[g_{\text{pass}}, 1]$ (通過域に於ける伝達関数の値域) の「適切な近傍」にあるものだけをすべて求める。そのようなベクトル \mathbf{v} は B -正規直交系にとれて、それらを集めて並べたものを組 Z にすればよい。

補足的考察

ここで区間 $[g_{\text{pass}}, 1]$ の「適切な近傍」とは、固有値 ϕ の分布を (誤差の影響を考慮した上で) はっきりと分離する近傍のことである。数値計算で求めた固有値 ϕ の値は近似の誤差や数値演算の丸め誤差、連立 1 次方程式の求解計算の不完全さなどに起因する誤差によって摂動を受ける。そのためたとえば g_{pass} よりも小さい ϕ の値でも、それが g_{pass} に非常に近ければ安全のためには含めておく必要がある。また、固有ベクトルの組が不変部分空間になるためには、縮退している固有値の固有ベクトルがある場合にはそれらの全部を含むかあるいは全く含まないかのどちらかであって半端に含んでいてはならないが、数値計算では本来は縮退しているはずの固有値が摂動を受けて縮退がほどけて分散している可能性があり、通常の数値計算では真に縮

退しているかを判断することができないので、可能であるならば摂動の限界を見積り、値の近接している固有値は本来は縮退している可能性があるためひとつの塊として扱う必要がある。実用的には結局、数値で求めた固有値 ϕ の値を g_{pass} 未満のものから (あるいは必要なものを落とさないように余裕を確保して g_{pass} よりも少しだけ小さい値、例えば $0.95g_{\text{pass}}$ などから開始して) 減少順に探索してゆき、 ϕ の分布の中に比較的大きな空隙 (すきま) のある場所を見つけて、それよりも上側にある ϕ の値はすべて近傍に含める、などとするのが安全であろう。伝達域での伝達率の区間の上限である 1 についても誤差の影響を考慮し、1 を越えている ϕ の値も近傍に含めるようにする。近傍に含めた最小の ϕ の値が、含めなかった最大の ϕ の値から十分に離れていて、しかも 1 に比べて微小でなければ、条件の良い近傍の設定に成功したことになる。もしも非常に微小な ϕ の値を近傍に含めてしまうと、あたかも g_{pass} を最初からそのような値に設定していたのと同様になってしまい、その後の計算で精度が低下する原因となるから、近傍に含める ϕ の値の大きさには下限を設定しておくのが良い。例えば $(1/2)g_{\text{pass}}$ よりも小さい ϕ は近傍に含めない、などとする。(ここで g_{pass} は、用いたフィルタの通過域に於ける伝達関数の「実際の」下限値である。今回作ったフィルタの伝達関数の g_{pass} は 3×10^{-6} や 1×10^{-4} のようにかなり小さい値なので、フィルタ対角化法だけで得られた近似解の精度はそれだけ落ちている可能性がある。図 13 や図 17 のグラフの線 (IT0) にその影響があらわれている。)

注: 固有対の抜け落ちをなるべく避けるためには、余裕をとるために区間 $[g_{\text{pass}}, 1]$ を少し広げてやり、固有値 ϕ の選別の閾値を下げて \mathcal{F} の近似固有対を多めに取り入れると、その結果として最終的に得られる (元の固有値問題の) フィルタ対角化法による近似固有対には、通過域 $[a, b]$ から少し離れた遷移域に固有値を持つものが含まれる可能性がある。そのような本来は不必要な固有対は、あとで固有値により選別を行えば取り除くことができる。固有値の値がきわどい位置にある近似対については、固有値による選別は精度を高める改良を加えた後で行う方が安全であろう。

A.1.2 \mathcal{F} の近似固有対の解法

ベクトル \mathbf{v} を Y の張る空間内に制限して得られる \mathcal{F} の近似固有対 (ϕ, \mathbf{v}) は、以下の方法で求めている。

固有値方程式 $\mathcal{F}\mathbf{v} = \phi\mathbf{v}$ に対して、 \mathbf{v} が Y の張る空間内にあること条件式 $\mathbf{v} = Y\mathbf{u}$ を課して、さらに両辺に左側から $X^T B$ を乗じると、次数 m の小規模な GEVP: $\alpha\mathbf{u} = \phi\beta\mathbf{u}$ が得られる (左側から任意のベクトルの組ではなくて $X^T B$ を乗じることによる利点は、それにより得られる小規模な GEVP の係数の行列 α, β が以下に示すような良い性質を持つことにある。) ここで $\alpha = X^T (B\mathcal{F})Y$, $\beta = X^T B Y$ である。いま行列 A, B は実対称で、また仮定から τ_p は

(A, B を係数とする一般固有値問題の固有値とは異なる) 実数であるから, $B\mathcal{F} = c_\infty B + B \sum_p (A - \tau_p B)^{-1} B$ により行列 $B\mathcal{F}$ は実対称である. このことから, まず $\alpha = X^T(B\mathcal{F})Y = X^T(B\mathcal{F})^T Y = X^T \mathcal{F}^T B Y = (\mathcal{F}X)^T B Y = Y^T B Y$ となるので行列 α は実対称半正定値であり, また $\beta = X^T B Y = X^T B \mathcal{F} X = X^T (B\mathcal{F})^T X = X^T \mathcal{F}^T B X = (\mathcal{F}X)^T B X = Y^T B X = \beta^T$ となるので行列 β は実対称であることがわかる. さらに, β の固有値はすべて $f(\lambda)$ の値域に含まれることが示せる (副節 A.1.4 に記述). (ただし, そのことから理論上は β が正定値や半正定値となる場合であっても, 数値計算で求めた固有対には微小な負の固有値を持つものが混入する. これは絶対値が小さい固有値を持つ固有対は数値誤差の影響で精度が出せないためである. また β が不定値の場合でも負の固有値がすべて微小であるならば, 数値計算としては実質的に半正定値と同様にみなせる. そこで後で見るように, 固有値の大きさが微小な β の固有ベクトルについての関係を切断処理により棄てる. それにより最終的な計算結果が大きく影響されない定式化になっていればそれで良い.)

フィルタの性質から自然に導かれることであるが, 通常は Y の特異値には (フィルタで濾過するベクトルの個数を十分大きくすればそれだけ) 微小な値が多くなり, 係数 α, β は共に悪条件になる. そのことを考慮して, 固有値方程式 $\alpha \mathbf{u} = \phi \beta \mathbf{u}$ は以下 (副節 A.1.3) で記述する方法により丁寧に解く.

A.1.3 「縮小された \mathcal{F} の固有値方程式」 $\alpha \mathbf{u} = \phi \beta \mathbf{u}$ の解法

行列 Y は一般には非常に悪条件となる可能性があるが, そのような行列 Y がそれぞれ α の表式中には 2 回, β には 1 回, 積の形で入っていることから, β の方が α よりも行列の数値的な条件が良いはずである. そこで実対称行列 β をまず対称行列用の Jacobi 対角化法を用いて直交変換で固有値分解する. これにより得られる固有値はすべて実数で, 固有ベクトルを固有値の値の減少順に対応して (負の固有値のものは末尾に置いて) 並べた β の固有値分解を $\beta = Q^T D Q$ とする

直交行列 Q による変換 $\mathbf{u} \equiv Q^T \mathbf{w}$ により方程式は $G \mathbf{w} = \phi D \mathbf{w}$ となる (但し, $G \equiv Q \alpha Q^T$ で, 行列 G は実対称である).

ここで数値計算上での安定化のために, 極めて小さい正の値 ϵ_{thres} を閾値として導入し, ϵ_{thres} 未満の β の固有値 (D の固有値) を切断する方法で「方程式の正則化」を行なう (たとえば $\epsilon_{\text{thres}} = 100 \epsilon_{\text{mac}}$ などとする. ここで ϵ_{mac} は使用する精度での浮動小数点数のマシンイプシロンを表す*)¹⁾. 閾値 ϵ_{thres} 未満である D の対角要素に対応する行

および列をそれぞれ G, D, \mathbf{w} から省くことで, 切断された固有値方程式 $\hat{G} \hat{\mathbf{w}} = \phi \hat{D} \hat{\mathbf{w}}$ が得られる (行列 \hat{G} は実対称で, \hat{D} は正の対角行列である). さらに, 対角スケール変換 $\hat{\mathbf{w}} \equiv \hat{D}^{-1/2} \mathbf{z}$ により $H \equiv \hat{D}^{-1/2} \hat{G} \hat{D}^{-1/2}$ とおいて得られる実対称行列 H の標準固有値問題 $H \mathbf{z} = \phi \mathbf{z}$ を再び対称行列用の Jacobi 対角化法を用いて解く.

その固有対 (ϕ, \mathbf{z}) のベクトル \mathbf{z} から逆変換 $\hat{\mathbf{w}} := \hat{D}^{-1/2} \mathbf{z}$ により対応する $\hat{\mathbf{w}}$ を得て, 切断された行の自由度に零を補って \mathbf{w} を得て, さらに $\mathbf{u} := Q^T \mathbf{w}$ とする.

このようにして構成された \mathbf{u} から $\mathbf{v} := Y \mathbf{u}$ で作った \mathbf{v} をすべて集めたものが, 元の一般固有値問題の不変部分空間を近似する空間の基底となる. ただし, この \mathbf{v} はそれだけではまだ B -正規直交系にはならない. いま上述のように, 実対称行列 H の標準固有値問題を Jacobi 対角化法を用いて解いたとすれば, 算法の性質から得られた固有ベクトル \mathbf{z} は自然に正規直交系になる. すると (β の微小な固有値を閾値で切断した影響を無視すれば) \mathbf{u} は β -正規直交系になることがわかる. そのときには上の $\mathbf{v}^{(i)} := Y \mathbf{u}^{(i)}$ の代わりにスカラ倍の因子を付け加えたもの $\mathbf{v}^{(i)} := (1/\sqrt{\phi^{(i)}}) \cdot Y \mathbf{u}^{(i)}$ を作ると, 数学的にはこの新しい $\mathbf{v}^{(i)}$ が B -正規直交系になることが示せる. ただし, この規格化因子 $(1/\sqrt{\phi^{(i)}})$ が示唆しているように, もしも通過域にある固有ベクトルの伝達率に相当する $\phi^{(i)}$ に小さい値があると数値計算では誤差が拡大して B -正規直交性が (規格化だけではなくて直交性も) 崩れる可能性がある. フィルタの伝達関数の g_{pass} の値が 1 の付近にある場合にはそのような困難は生じないが, 今回用いたフィルタのように伝達関数の特性値 g_{pass} が 3×10^{-6} や 1×10^{-4} などの比較的小さい値を持つ特性の良くないフィルタを用いた場合には, 数値丸め誤差の拡大によって B -正規直交性が崩れる可能性があるので, 今回の実験ではそのような場合の崩れを補正する B -正規直交化を追加している.

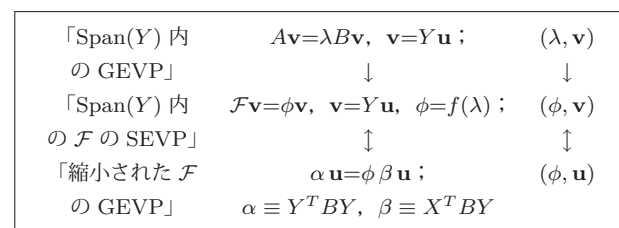


図 A.1 対応関係の図

Fig. A.1 Illustration of correspondences

A.1.4 付記：行列 β の固有値の範囲

フィルタ作用素の固有値問題 $\mathcal{F} \mathbf{w} = \phi \mathbf{w}$ と等価な実対称定値一般固有値問題 $(B\mathcal{F}) \mathbf{w} = \phi B \mathbf{w}$ を考え, それに対応する (一般) Rayleigh 商 $Q = \frac{\mathbf{w}^T (B\mathcal{F}) \mathbf{w}}{\mathbf{w}^T B \mathbf{w}}$ (但し \mathbf{w} は非

であれば約 2.22×10^{-16} である. Fortran90 言語では組込関数 EPSILON を用いることで簡単に表せる.

*1) マシンイプシロン ϵ_{mac} は 1 よりも大きい最小の浮動小数点数値と 1 との差で, たとえば IEEE 754 規格の 2 進 64bit 倍精度

零の実ベクトル) を作ると, Q の最大値と最小値はそれぞれ伝達率 $\phi^{(i)}$ の最大値と最小値である.

いま非零の実ベクトル \mathbf{w} のとりうる範囲を X の各列で張られた線形部分空間に制限する条件 $\mathbf{w} = X\mathbf{u}$ (但し \mathbf{u} は任意の非零の実ベクトル) を Q に課して得られる値を Q' とすると, $X^T B F X = X^T B Y = \beta$, $X^T B X = I_m$ であることから $Q' = \frac{\mathbf{u}^T \beta \mathbf{u}}{\mathbf{u}^T \mathbf{u}}$ とかけるが, これは実対称行列 β の標準固有値問題に対応する Rayleigh 商なので, Q' の最大値と最小値はそれぞれ β の固有値の最大値と最小値になる. すると制限を課した後の Q' の値域は制限を課す前の Q の値域に含まれることから, 伝達率 $\phi^{(i)}$ の最大値と最小値によりそれぞれ β の固有値の上限と下限が与えられることがわかる. あるいは元の GEVP の固有値の下限を a とするとき, すべての β の固有値は伝達関数 $f(\lambda)$ の $a \leq \lambda$ に於ける値域に含まれることがわかる.

参考文献

- [1] 村上 弘: 固有値が指定された区間にある固有対を解くための対称固有値問題用のフィルタの設計, **情報処理学会論文誌: コンピューティングシステム** (ACS31), Vol.3, No.3, pp.1-21 (2010).
- [2] 村上 弘: フィルタで濾過されたベクトルの組から不変部分空間の直交基底の組を近似構成するフィルタ対角化法, **情報処理学会研究報告**, Vol.2011-HPC-129, No.1, pp.1-8 (2011).
- [3] 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** (ACS35), Vol.4, No.4, pp.1-14 (2011).
- [4] 村上 弘: 実対称定値一般固有値問題のフィルタ対角化による解法の数理, **情報処理学会研究報告**, Vol.2011-HPC-131, No.5, pp.1-12 (2011).
- [5] 村上 弘: レゾルベントを用いたフィルタによる固有値問題の解法について, **情報処理学会研究報告**, Vol.2012-HPC-133, No.22, pp.1-8 (2012).
- [6] 村上 弘: レゾルベントの線形結合によるフィルタの伝達特性の調整, **京都大学数理解析研究所講究録**, No.1785, pp.131-145 (2012).
- [7] 村上 弘: レゾルベントの線形結合をフィルタに用いたエルミート定値一般固有値問題のフィルタ対角化法, **情報処理学会論文誌: コンピューティングシステム** (ACS45), Vol.7, No.1 (2014年3月), pp.57-72.
- [8] Anthony P. Austin and Lloyd N. Trefethen: Computing eigenvalues of real symmetric matrices with rational filters in real arithmetic, *preprint submitted to* , **SIAM J. Sci. Comput.** (2014).