Filter diagonalization method for a real symmetric definite GEVP whose filter is a polynomial of a resolvent

Hiroshi Murakami (Tokyo Metropolitan University)

ABSTRACT

For a real-sym-def gen-EVP $Av = \lambda Bv$ $(B > 0)$, we solve eigenpairs whose eigenvalues are in a neighbor of a real interval $[a, b]$ by the filter diagonalization method.

In our current study, the filter is the real part of a polynomial of a resolvent: $\mathcal{F} = \mathrm{Re}\sum_{k=1}^{n} \gamma_k \{\mathcal{R}(\rho)\}^k$. Here $\mathcal{R}(\rho) \equiv (A - \rho B)^{-1}B$ is the resolvent with an imaginary shift ρ, and γ_k are coefficients. In our experiments, the (half) degree n is 15 or 20.

The shift ρ and the set of coefficients $\{\gamma_k\}$ of the filter are tuned so that the filter passes those eigenvectors well whose eigenvalues are in a neighbor of $[a, b]$ but reduces those eigenvectors strongly whose eigenvalues are separated from the interval.

An application of the filter to a set of sufficiently many B-orthonormal random vectors $\{x^{(\ell)}\}$ gives another set $\{y^{(\ell)}\}$. From both sets of vectors and also properties of the filter, a basis is constructed which spans an approximation of the invariant subspace whose eigenvalues are in a neighbor of $[a, b]$. An application of Rayleigh-Ritz procedure to the basis gives approximations of all required eigenpairs.

Experiments for banded problems showed this approach worked in success.

Introduction

- In previous studies, the filter we used was a linear combination of resolvents. For example, 6 to 16 resolvents were used.
- The action of the resolvent $\mathcal{R}(\rho) \equiv (A - \rho B)^{-1}B$ is, a multiplication of B and the solution of LEQ whose coefficient is $C \equiv A - \rho B$.
  - When C is banded, the LEQ is solved by some direct method such as LU factorization.
  - When C is random sparse, the LEQ is solved by some iterative method using incomplete LU factorization.
- In application of the filter, matrix factorization is the large portion.
- The amount of memory to keep the matrix factor is a severe constraint in large size calculation.

  If we use many resolvents and they are applied in parallel, the total amount of memory required is proportional to the number of resolvents.

The present approach

- We assume the amount of memory is the severest constraint.

  In present study, we try a method which uses only one resolvent rather than many.
- The filter we use is a *polynomial* of a resolvent.

  In the application of the filter,
  the resolvent is applied n times when the polynomial is degree n.
- The action of a resolvent is by the solution of a LEQ which uses factorization of the matrix.

  The factor is made once and kept,
  and it is used when the resolvent is applied.

Filter and its transfer function

We consider a real sym. def. GEVP $A\,v = \lambda\,B\,v$, where B is pos.def..

- $\mathcal{R}(\rho) \equiv (A - \rho B)^{-1}B$ is the resolvent with shift ρ.

  For any eigenpair (λ, v), we have $\mathcal{R}(\rho)\,v = \dfrac{1}{\lambda - \rho}\,v$.
- Filter $\mathcal{F}$ is the real part of a (half) degree n polynomial of $\mathcal{R}(\rho)$:

$$\mathcal{F} = c_\infty\,I + \mathrm{Re}\sum_{k=1}^{n} \gamma_k \{\mathcal{R}(\rho)\}^k\,.$$

  For any eigenpair (λ, v), we have $\mathcal{F}\,v = f(\lambda)\,v$.
- Here, $f(\lambda)$ is the transfer function of the filter $\mathcal{F}$ which is a real rational function of λ whose poles are only at ρ and its complex conjugate $\bar{\rho}$.

$$f(\lambda) = c_\infty + \mathrm{Re}\sum_{k=1}^{n} \frac{\gamma_k}{(\lambda - \rho)^k}$$

- We are to solve those eigenpairs whose eigenvalues are in $[a, b]$.

  The normalized coordinate t of λ is defined by the linear transformation $\lambda = \dfrac{a+b}{2} + t\,\dfrac{b-a}{2}$ which maps between $\lambda \in [a, b]$ and $t \in [-1, 1]$.
  - passband : $t \in [-1, 1]$
  - transition region : $1 < |t| < \mu$
  - stopbands : $\mu \le |t|$
- Transfer function in normalized coordinate t is $g(t) \equiv f(\lambda)$.

  $g(t) \ge g_{\rm pass}$ if and only if t is in the passband.

  $|g(t)| \le g_{\rm stop}$ if t is in stopbands.
- We restrict $g(t)$ to an even function, then two poles are a pair of complex conjugates and pure imaginary numbers.

- We place the pure imaginary poles of $g(t)$ at $t = \pm\sqrt{-1}$:

$$g(t) = c'_\infty + \mathrm{Re}\sum_{k=1}^{n} \frac{\alpha_k}{(1 + t\sqrt{-1})^k}\,.$$

- Coefficients α_k, k=$1, 2, \ldots, n$ are real numbers, to make $g(t)$ an even function.
- The real parameters α_k, k=$1, 2, \ldots, n$ are tuned :
  - In the passband $|t| \le 1$, the value of $g(t)$ is close to 1.
  - In stopbands $\mu \le |t|$, the magnitude of $g(t)$ is very small.

  In present study, the parameters are optimized by a LSQ-like method.

Reverse construction of $\mathcal{F}$ from $g(t)$

From $g(t)$ we construct the filter operator $\mathcal{F}$. Since:

$$g(t) \;= c'_\infty + \mathrm{Re}\sum_{k=1}^{n} \frac{\alpha_k}{(1 + t\sqrt{-1})^k}\,,$$

$$f(\lambda) \;= c_\infty + \mathrm{Re}\sum_{k=1}^{n} \frac{\gamma_k}{(\lambda - \rho)^k}\,,$$

and also the relations $f(\lambda) = g(t)$ and $\lambda = \dfrac{a+b}{2} + t\,\dfrac{b-a}{2}$, we have

$$c'_\infty = c_\infty, \quad \gamma_k = \left(-\frac{b-a}{2}\sqrt{-1}\right)^k \alpha_k,\ \ k{=}1, 2, \ldots, n,$$

$$\rho = \frac{a+b}{2} + \frac{b-a}{2}\sqrt{-1}.$$

For simplicity, the transfer rate at infinity c_∞ is set to zero.

$$\mathcal{F} \;=\; \mathrm{Re}\sum_{k=1}^{n} \gamma_k \{\mathcal{R}(\rho)\}^k\,.$$

Calculation of the action of the filter $\mathcal{F}$

The filter is specified by degree n, shift $\rho \in \mathbb{C}$ and coefs $\gamma_k \in \mathbb{C}$, k=$1, 2, \ldots, n$:

$$\mathcal{F} \;=\; \mathrm{Re}\sum_{k=1}^{n} \gamma_k \{\mathcal{R}(\rho)\}^k\,.$$

Let X and Y are real $N{\times}m$ matrices which are sets of m real column vectors of size N.

Then the action of degree n filter $Y \leftarrow \mathcal{F}\,X$ is calculated by:

```
W ← X ;
Y ← 0 ;
for k := 1 to n do begin
   Z ← R(ρ) W ;
   Y ← Y + Re(γk Z) ;
   W ← Z
end
```

Here, W and Z are complex $N{\times}m$ matrices (just for work).

Calculation of the action of a resolvent

- To calculate the action of the resolvent $Z \leftarrow \mathcal{R}(\rho)\,W$,
  first the r.h.s. $B\,W$ is calculated from W,
  then the LEQ $C\,Z = B\,W$ is solved for Z
  whose coefficient matrix is $C \equiv A - \rho B$.
- Since both matrices A and B are real symmetric,
  C is complex symmetric ($C^T = C$).
  When both matrices A and B are banded, C is also banded.
- In present experiments,
  the complex modified Cholesky factorization is used
  for the complex banded symmetric matrix C.

Filters used in experiments

Coefficients α_k, k=$1, 2, \ldots, n$ of the filters from (no.1) to (no.3) are obtained by LSQ-like method.

- Filter (no.1) : (half) degree $n = 15$.

  $\mu = 2.0$, $g_{\rm pass} = 2.3{\times}10^{-4}$, $g_{\rm stop} = 1.1{\times}10^{-15}$.
- Filter (no.2) : (half) degree $n = 15$.

  $\mu = 1.5$, $g_{\rm pass} = 5.46{\times}10^{-5}$, $g_{\rm stop} = 5.85{\times}10^{-13}$.

  μ is set smaller than the case of filter (no.1).
  In exchange, $g_{\rm pass}$ is smaller and $g_{\rm stop}$ is larger.
- Filter (no.3) : (half) degree $n = 20$.

  $\mu = 2.0$, $g_{\rm pass} = 1.273{\times}10^{-2}$, $g_{\rm stop} = 2.6{\times}10^{-15}$.

  By choosing higher n, the increased degrees of freedom makes $g_{\rm pass}$ closer to 1 than the case of filter (no.1).

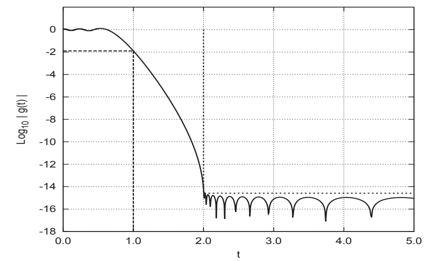

Fig 1. Filter (no.3): magnitude of transfer func $|g(t)|$

Test problem : EVP of 3D-Laplacian discretized by FEM

- The test problem is EVP of Laplacian in 3D region $[0, \pi]{\times}[0, \pi]{\times}[0, \pi]$ with zero-Dirichlet boundary condition :

$$-\nabla^2\,\Psi(x, y, z) = \lambda\,\Psi(x, y, z)\,.$$

  FEM discretization gives a real symmetric definite GEVP:

$$A\,v = \lambda B\,v\,.$$

  We solve eigenpairs (λ, v) whose eigenvalues are in a specified interval $[a, b]$.

- Each direction of the edge of cubic region is equi-divided into n_1+$1 = 51$, n_2+$1 = 61$, n_3+$1 = 71$ sub-intervals to make finite elements.

  Basis functions inside each FEM element are products of piece-wise linear function in each direction.
- Size of both matrices A and B is $N = n_1\,n_2\,n_3 = 50 \times 60 \times 70 = 210,000$.
- By a good numbering of basis functions, the lower bandwidth of matrices is $1 + n_1 + n_1 n_2 = 1 + 50 + 50 \times 60 = 3,051$

  (Although A and B are quite sparse inside their bands,
  in the calculation they are treated as if dense).
- Eigenpairs are solved whose eigenvalues are in $[200, 210]$

  (True count of eigenpairs is 91).

  By this discretization, no eigenvalues are degenerate
  since sub-divisions are made differently in each direction.
- For this problem, exact eigenvalues can be calculated by a formula.

  We found errors of approximated eigenvalues are less than 4×10^{-13}.

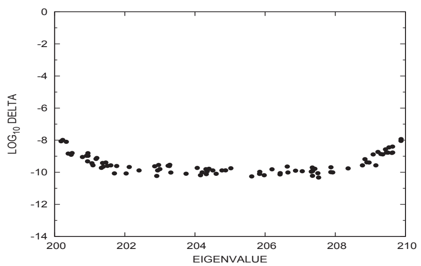

Fig 2. Filter (no.3) : Δ the B^{-1}-norm of residual of approximated pairs. ($[200, 210]$, m=200)

Elapse times to solve Eigenpairs

Size of matrices : $N = 50 \times 60 \times 70 = 210,000$.
Lower bandwidth : $w_L = 1 + 50 + 50 \times 60 = 3,051$.
Number of vectors filtered : $m = 200$.
Interval of eigenvalue to solve : $[a,\ b] = [200,\ 210]$.
The true count of eigenvalues : 91.
Method of matrix factorization : complex banded mod-Cholesky.
Machine: intel Corei7-5960X with 64GB mem(8core, HT=off, TB=off).

Kind of filter	(no.1)	(no.2)	(no.3)
Whole F-D-M (in sec)	2,490.3	2,491.2	3,105.0
−− Generation of random vectors	0.2	0.2	0.2
−− B-orthonormalization of random vecs	82.3	82.3	82.4
−− Application of the filter	2,140.7	2,141.6	2,755.2
−− Construction of basis of inv-subspace	193.5	193.4	193.5
−− Rayleigh-Ritz	73.6	73.7	73.7
Memory usage (GB)(virtual,real)	21.5(20)	21.5(20)	21.5(20)

Conclusion

- Filter diagonalization method solves eigenpairs of GEVP whose eigenvalues are in the specified interval. In present study, we used a filter which is a *polynomial* of a resolvent to reduce the amount of required memory and computation.
- In present experiments, the degree of the polynomial is n=15 or n=20, and the set of coefficients of the polynomial is determined by a LSQ-like method.
- Compared from the case when the filter is a linear combination of many resolvents, only one resolvent is required. The resolvent is applied n times during the filtering process, therefore the amount of computation can be reduced if the matrix decomposition is made once and it is used n times to make the actions of the resolvent.
- We made numerical experiments and obtained consistent results.