2020 年 2 月 计 算 数 学 第 42 卷第 1 期
Feb., 2020 MATHEMATICA NUMERICA SINICA Vol.42, No.1

计算矩阵函数双线性形式的 Krylov 子空间算法的误差分析*1)

贾仲孝 孙晓琳
(清华大学数学科学系, 北京 100084)

摘 要

矩阵函数的双线性形式 $u^Tf(A)v$ 出现在很多应用问题中, 其中 $u,v\in\mathbb{R}^n, A\in\mathbb{R}^{n\times n}$, $f(z)$ 为给定的解析函数. 开发其有效可靠的数值算法一直是近年来学术界所关注的问题, 其中关于其数值算法的停机准则多种多样, 但欠缺理论支持, 可靠性存疑. 本文将对矩阵函数的双线性形式 $u^Tf(A)v$ 的数值算法和后验误差估计进行研究, 给出其基于 Krylov 子空间算法的误差分析, 导出相应的误差展开式, 证明误差展开式的首项是一个可靠的后验误差估计, 据此可以为算法设计出可靠的停机准则.

关键词: 双线性形式; Krylov 子空间方法; 相对误差估计; 停机准则
MR (2010) 主题分类: 65F30, 65F10

1. 引 言

本文研究矩阵函数双线性形式的数值计算方法及其后验误差估计, 具体的数学表达式如下:

$$u^Tf(A)v, \tag{1.1}$$

矩阵 $A\in\mathbb{R}^{n\times n}$ 是大规模稀疏矩阵, u, v 满足 $\|u\|=1$, $\|v\|=1$, 这里 $\|\cdot\|$ 是 2 范数, 上标 T 表示向量或矩阵的转置, $f(z)$ 是在给定区域上的解析函数, 且使得矩阵函数 $f(A)$ 有定义. 我们主要对实矩阵进行研究, 复矩阵除了计算复杂性较实矩阵高之外, 理论上没有本质区别.

矩阵函数双线性形式 $u^Tf(A)v$ 的应用十分广泛, 如积分方程的数值求解、模型降阶 [14]、共轭梯度法的计算精度估计 [10] 等等. 关于其算法的分析与设计, Gene Golub [8–10] 是其关键人物. 他将双线性形式 (1.1) 化成一种 Riemann-Stieltjes 积分 [5–7, 9] 的形式, 而后采用求积公式得到了数值计算矩阵函数双线性形式的一个近似值, 这也是近年来求解矩阵函数二次型问题最常采用的方法 [11]. 但是关于该方法尚无一个可设计为停机准则的后验误差估计 [12, 13]. 但是我们发现其积分近似值本质上与采用 Krylov 子空间方法计算 $u^Tf(A)v$ 得到的结果一致, 而针对给定区域内的解析函数, 关于计算矩阵函数乘向量 $f(A)v$ 的 Krylov 子空间方法 [2] 已有完备的后验误差估计理论. 因此, 从 Krylov 子空间的角度出发, 我们可以将数值计算 $f(A)v$ 的 Krylov 子空间方法的后验误差估计理论推广到矩阵函数双线性形式 $u^Tf(A)v$ 上, 从而可以建立可靠的后验误差估计, 以作为停机准则.

本文首先给出计算 $u^Tf(A)v$ 的 Krylov 子空间方法, 并给出其误差估计的分析. 当矩阵函数在给定区域内解析并满足一定条件时, 给出关于其误差的误差展开式, 理论上证明误差展开

* 2018 年 10 月 5 日收到.
1) 基金项目: 国家自然科学基金资助 (项目编号 11771249).

式的首项是一个可靠后验误差估计, 最后报告数值实验, 针对一些常用的函数, 如指数函数、三角函数, 验证误差估计的可靠性. 对一些非解析函数, 如双曲函数, 虽然不能理论上证明误差展开式首项是一个可靠的后验误差估计, 但数值实验表明, 所导出的误差估计仍然是真实误差的可靠估计.

文中 $\mathbb{C}^{n\times n}$ 或 $\mathbb{R}^{n\times n}$ 表示 n 阶复矩阵或实矩阵集合, $\mathbb{C}^n$ 或 $\mathbb{R}^n$ 表示 n 维复列向量或实向量集合, A^H 或 A^T 表示矩阵 A 的共轭转置或转置. $||\cdot||$ 表示向量或矩阵的 2- 范数, $||\cdot||_F$ 表示矩阵的 F- 范数. $spec(A)$ 表示矩阵 A 的谱, $\mathcal{F}(A)$ 表示矩阵 A 的数值域, 即 $\mathcal{F}(A)=\{x^HAx: x\in\mathbb{C}^N, x^Hx=1\}$. $\lambda_{\max}(A)$ 和 $\lambda_{\min}(A)$ 表示实对称矩阵 A 的最大和最小特征值, I 表示相应阶数的单位矩阵, e_j 表示单位矩阵的第 j 列.

2. 计算 $u^Tf(A)v$ 的 Krylov 子空间方法

矩阵函数的双线性形式 $u^Tf(A)v$ 涉及到矩阵函数 $f(A)$ 和矩阵函数乘向量 $f(A)v$, 计算矩阵函数的标准算法均没有考虑到矩阵 A 可能有的特殊结构, 在计算过程中可能破坏其特殊结构, 因此在针对具有特殊结构的大规模矩阵会导致计算量和存储量过大, 不能接受. 所以问题的关键还是计算矩阵函数乘向量 $f(A)v$, 对于大规模矩阵, 最经典方法的是基于 Krylov 子空间的方法. 在此基础上, 想要得到矩阵函数的双线性形式只需要再进行一步向量内积即可. 计算矩阵函数乘向量的 Krylov 子空间算法本质上属于投影类算法, 投影类算法在求解大规模线性代数问题时十分有效, 因此科研工作者在研究矩阵函数乘向量的有效数值算法中开发出了相应的投影类算法, 其投影空间是 Krylov 子空间. 该方法可以充分利用矩阵 A 的稀疏性, 因为在形成子空间的一组基底时只利用到了矩阵向量乘的操作. Krylov 子空间算法将原大规模问题 $f(A)v$ 投影到了一个低维的子空间, 然后通过标准算法来求解相应的中小规模的问题即可得到其近似.

这里我们采取 Arnoldi 方法来近似 $u^Tf(A)v$, 其 Arnoldi 近似基于如下的 Arnoldi 分解:

$$AV_k = V_kH_k + h_{k+1,k}v_{k+1}e_k^T, \tag{2.1}$$

其中 $V_k=[v_1,v_2,\cdots,v_k]$ 的列构成了 Krylov 子空间 $\mathcal{K}_k(A,v_1)=\text{span}\{v_1,Av_1,\cdots,A^{k-1}v_1\}$ 的一组标准正交基, $H_k=[h_{i,j}]$ 为 k 阶上 Hessenberg 矩阵, $e_k\in\mathbb{R}^k$ 为 k 阶单位阵的第 k 列.

由 (2.1) 可得 $H_k=V_k^TAV_k$, 因此有

$$u^Tf(A)v\approx\beta u^TV_kV_k^Tf(A)V_ke_1\approx\beta u^TV_kf(V_k^TAV_k)e_1=\beta u^TV_kf(H_k)e_1\equiv F_k, \tag{2.2}$$

F_k 即为 $u^Tf(A)v$ 的 Arnoldi 近似, 上述方法称为标准 Arnoldi 方法, 其中 $\beta=\|v\|$, Arnoldi 分解中的初始向量 $v_1=v/\beta$.

3. $E_k(f)$ 的误差展开式

令

$$F_k=\beta u^TV_kf(H_k)e_1 \tag{3.1}$$

为 $u^Tf(A)v$ 的 Arnoldi 近似. 定义误差

$$E_k(f)=u^Tf(A)v-F_k=u^Tf(A)v-\beta u^TV_kf(H_k)e_1. \tag{3.2}$$

本文假设函数 $f(z)$ 在包含矩阵 A 和 H_k 的数值域 $\mathcal{F}(A)$ 和 $\mathcal{F}(H_k)$ 的闭凸集 S 及其边界上解析. 点列 $\{t_k\}_{k=0}^{\infty}$ 来自 S, 且其中相同点连续排列, 即 $t_i = t_{i+1} = \cdots = t_j$, 函数 $\phi_j(t)$ 是满足如下条件的解析函数列:

$$\begin{cases} \phi_0(t) = f(t) \\ \phi_{j+1}(t) = \frac{\phi_j(t)-\phi_j(t_j)}{t-t_j} \quad j \geq 0. \end{cases} \tag{3.3}$$

下面的结果成立.

定理 1. 假定 $f(z)$ 在包含矩阵 A 和 H_k 的数值域 $\mathcal{F}(A)$ 和 $\mathcal{F}(H_k)$ 的闭凸集 S 及其边界上解析, 存在常数 C 使得对任何 $j \geq 0$ 有 $\max_{z\in S}|f^{(j)}(z)| \leq C$, 则对于 S 中给定的点列 $\{t_k\}_{k=0}^{\infty}$, $u^T f(A)v$ 基于 Arnoldi 近似的误差展开式如下:

$$E_k(f) = u^T f(A)v - \beta u^T V_k f(H_k)e_1 = \beta h_{k+1,k} \sum_{j=1}^{\infty} e_k^T \phi_j(H_k) e_1 u^T p_{j-1}(A) v_{k+1}. \tag{3.4}$$

其中 $p_0(t) = 1$, $p_j(t) = (t-t_0)(t-t_1)\cdots(t-t_{j-1})$, $j \geq 1$, $t_i \in S$, $i \geq 0$.

证明. 首先定义 $\phi_i(A)v$ 的 Arnoldi 近似误差如下:

$$s_k^i = \phi_i(A)v - \beta V_k \phi_i(H_k)e_1. \tag{3.5}$$

由递归关系式 (3.3) 可得 $f(t) = f(t_0) + (t-t_0)\phi_1(t)$. 利用 Arnoldi 过程 (2.1), 有

$$\begin{aligned} u^T f(A)v &= f(t_0)u^T v + u^T(A-t_0 I)\phi_1(A)v \\ &= f(t_0)u^T v + u^T(A-t_0 I)(\beta V_k \phi_1(H_k)e_1 + s_k^1) \\ &= f(t_0)u^T v + \beta u^T(V_k(H_k - t_0 I) + h_{k+1,k}v_{k+1}e_k^T)\phi_1(H_k)e_1 + u^T(A-t_0 I)s_k^1 \\ &= \beta u^T V_k(f(t_0)e_1 + (H_k - t_0 I)\phi_1(H_k)e_1) \\ &\quad + \beta h_{k+1,k}u^T v_{k+1}e_k^T \phi_1(H_k)e_1 + u^T(A-t_0 I)s_k^1 \\ &= \beta u^T V_k f(H_k)e_1 + \beta h_{k+1,k}u^T v_{k+1}e_k^T\phi_1(H_k)e_1 + u^T(A-t_0 I)s_k^1. \end{aligned} \tag{3.6}$$

类似地, 对于 $\phi_1(A)v$ 有

$$\begin{aligned} \phi_1(A)v &= \phi_1(t_1)v + (A-t_1 I)\phi_2(A)v \\ &= \phi_1(t_1)v + (A-t_1 I)(\beta V_k \phi_2(H_k)e_1 + s_k^2) \\ &= \phi_1(t_1)v + \beta(V_k(H_k - t_1 I) + h_{k+1,k}v_{k+1}e_k^T)\phi_2(H_k)e_1 + (A-t_1 I)s_k^2 \\ &= \beta V_k(\phi_1(t_1)e_1 + (H_k - t_1 I)\phi_2(H_k)e_1) \\ &\quad + \beta h_{k+1,k}v_{k+1}e_k^T\phi_2(H_k)e_1 + (A-t_1 I)s_k^2 \\ &= \beta V_k \phi_1(H_k)e_1 + \beta h_{k+1,k}v_{k+1}e_k^T\phi_2(H_k)e_1 + (A-t_1 I)s_k^2. \end{aligned}$$

又因为 $s_k^i = \phi_i(A)v - \beta V_k\phi_i(H_k)e_1$, 取 $i=1$ 并将上述结果带入可得:

$$s_k^1 = \phi_1(A)v - \beta V_k\phi_1(H_k)e_1 = \beta h_{k+1,k}v_{k+1}e_k^T\phi_2(H_k)e_1 + (A-t_1 I)s_k^2.$$

同理, 对于 s_k^2, s_k^3, $\cdots$ 有类似形式结果成立, 即其通项有如下形式:

$$s_k^{i-1} = \beta h_{k+1,k}v_{k+1}e_k^T\phi_i(H_k)e_1 + (A-t_{i-1}I)s_k^i, \quad i = 2,3,\cdots,\infty,$$

将上述通项取 $i=1,2,\cdots$ 带入 (3.6) 式, 可得:

$$\begin{aligned} E_k(f) &= u^T f(A)v - \beta u^T V_k f(H_k)e_1 \\ &= \beta h_{k+1,k}\sum_{j=1}^{i} e_k^T \phi_j(H_k)e_1 u^T p_{j-1}(A)v_{k+1} + u^T p_i(A)s_k^i. \end{aligned}$$

下面给出关于 $|u^T p_i(A)s_k^i|$ 收敛性 (速度) 的分析. 证明存在正整数 P, 当 $i \geq P$ 时, $|u^T p_i(A)s_k^i|$ 收敛于 0 的速度快于 $1/i^{3/2}$.

由函数列 $\{\phi_j(t)\}$ 的定义可知, $\phi_{j+1}(t)$ 可以用函数差商来表示, 如下

$$\phi_{j+1}(t) = f[t, t_0, t_1, \cdots, t_j]. \tag{3.7}$$

因为 $f(z)$ 在 S 上 j 阶可导, 所以存在 $\zeta \in S$ 使得

$$f[t_0, t_1, \cdots, t_j] = \frac{f^{(j)}(\zeta)}{j!}. \tag{3.8}$$

因此, 由 $\max_{z\in S}|f^{(j)}(z)| \leq C$ 可得

$$|\phi_i(t)| = |f[t, t_0, \cdots, t_{i-1}]| \leq \frac{C}{i!}. \tag{3.9}$$

设矩阵 A 和矩阵 H_k 的 Schur 分解分别为: $Q_1^H A Q_1 = U_1$ 和 $Q_2^H H_k Q_2 = U_2$, 其中 Q_1, Q_2 分别为酉阵, U_1, U_2 为上三角矩阵. 记 $\hat{U}_1, \hat{U}_2$ 分别为 U_1, U_2 的严格上三角部分, 则关于 $\phi_i(A)$ 和 $\phi_i(H_k)$ 分别有下式成立

$$||\phi_i(A)|| \leq \sum_{m=0}^{n-1} \sup_{z\in S}|\phi_i^{(m)}(z)|\frac{||\hat{U}_1||_F^m}{m!}, ||\phi_i(H_k)|| \leq \sum_{m=0}^{k-1} \sup_{z\in S}|\phi_i^{(m)}(z)|\frac{||\hat{U}_2||_F^m}{m!}. \tag{3.10}$$

关于 $||\phi_i(A)||$ 和 $||\phi_i(H_k)||$ 的上界估计式可参考 [1] 中定理 2.4. 由式 (3.9) 和式 (3.10) 可得

$$||\phi_i(A)|| \leq \frac{C}{i!}\sum_{m=0}^{n-1}\frac{||\hat{U}_1||_F^m}{m!}, ||\phi_i(H_k)|| \leq \frac{C}{i!}\sum_{m=0}^{k-1}\frac{||\hat{U}_2||_F^m}{m!}. \tag{3.11}$$

由 $s_k^i = \phi_i(A)v - \beta V_k \phi_i(H_k)e_1$ 和上面的结果可得

$$\begin{aligned} ||s_k^i|| &\leq ||\phi_i(A)||||v|| + \beta||V_k||||\phi_i(H_k)||||e_1|| \\ &\leq \frac{1}{i!}(C\beta\sum_{m=0}^{n-1}\frac{||\hat{U}_1||_F^m}{m!} + C\beta||V_k||\sum_{m=0}^{k-1}\frac{||\hat{U}_2||_F^m}{m!}) \\ &= \frac{C_1}{i!}. \end{aligned} \tag{3.12}$$

又由于 $\{t_j\}_{k=0}^{i-1}$ 属于闭凸集 S, 因此

$$\begin{aligned} ||p_i(A)|| &= ||(A-t_0 I)(A-t_1 I)\cdots(A-t_{i-1}I)|| \\ &\leq \prod_{j=0}^{i-1}(||A|| + |t_j|) \leq C_2^i, \end{aligned} \tag{3.13}$$

其中 $C_2 = \|A\| + T$, T 表示闭凸集 S 中元素的最大绝对值. 结合 (3.12)、(3.13) 和 Cauchy-Schwarz 不等式, 有

$$|u^T p_i(A)s_k^i| \le ||u||||p_i(A)s_k^i|| \le ||u||C_1\frac{C_2^i}{i!} = C_3\frac{C_2^i}{i!}, \tag{3.14}$$

其中 $C_3 = ||u||C_1$, 再结合 Stirling 不等式 (参考 [4] 的第 257 页)

$$\sqrt{2\pi i}\left(\frac{i}{e}\right)^i < i! < \sqrt{2\pi i}\left(\frac{i}{e}\right)^i e^{\frac{1}{12i}}, \tag{3.15}$$

其中 e 是自然指数, 则 (3.14) 式可以继续放大为

$$|u^T p_i(A)s_k^i| \le C_3\frac{C_2^i}{i!} \le \frac{C_3}{\sqrt{2\pi i}}\left(\frac{C_2 e}{i}\right)^i. \tag{3.16}$$

对 $\left(\frac{C_2 e}{i}\right)^i$ 而言, 存在正整数 P , 当 $i \ge P$ 时, $\left(\frac{C_2 e}{i}\right)^i$ 收敛于 0 的速度要快于 $1/i$, 因此 $|u^T p_i(A)s_k^i|$ 趋于 0, 且当 $i \ge P$ 时, 其收敛速度快于 $1/i^{3/2}$. 因此, $u^T f(A)v$ 基于 Arnoldi 近似的误差展开式如下:

$$E_k(f) = u^T f(A)v - \beta u^T V_k f(H_k)e_1 = \beta h_{k+1,k}\sum_{j=1}^{\infty} e_k^T\phi_j(H_k)e_1 u^T p_{j-1}(A)v_{k+1},$$

即 (3.4) 成立, 定理得证.

注意到, 如果点列 $\{t_i\}_{i=0}^{n-1}$ 是矩阵 A 的 n 个特征值, 则误差展开式 (3.4) 可化为有限项的和:

$$E_k(f) = u^T f(A)v - \beta u^T V_k f(H_k)e_1 = \beta h_{k+1,k}\sum_{j=1}^{n-1} e_k^T\phi_j(H_k)e_1 u^T p_{j-1}(A)v_{k+1}. \tag{3.17}$$

注 1. 理论表明, 存在正整数 P, 使得当 $i \ge P$ 时, $\left(\frac{C_2 e}{i}\right)^i$ 趋于 0 的速度要快于 $\frac{1}{i}$ 趋于 0 的速度, 也就是 $|u^T p_i(A)s_k^i|$ 收敛于 0 的速度大于 $1/i^{3/2}$. 为了寻找满足上述条件的正整数 P, 需要满足不等式

$$\left(\frac{C_2 e}{i}\right)^i < \frac{1}{i}, \tag{3.18}$$

即

$$\left(\frac{i}{C_2 e}\right)^i > i. \tag{3.19}$$

两边同时取对数并整理得

$$\log i > \log(C_2 e) + \frac{\log i}{i}. \tag{3.20}$$

当 $i \ge 1$ 时, $0 \le \frac{\log i}{i} < 1$, 因此放大上述不等式可得

$$\log i > \log(C_2 e) + 1 = \log(10C_2 e). \tag{3.21}$$

取 $P = \lceil 10C_2 e\rceil$, 则当 $i \ge P$ 时, 有

$$|\beta h_{k+1,k}\sum_{j=P}^{\infty} e_k^T\phi_j(H_k)e_1 u^T p_{j-1}(A)v_{k+1}| \le \sum_{P}^{\infty} i^{-3/2} \le \int_{P-1}^{\infty} x^{-3/2}dx = \frac{2}{\sqrt{P-1}}. \tag{3.22}$$

(3.16) 式意味着 (3.4) 的前 $P-1$ 项每项大小与 $1/i^{1/2}$ 同阶量. 再由上式可知, 通常情况下误差展开式的第一项的绝对值与误差展开式的绝对值同阶, 因此误差展开式的第一项的绝对值大小可以作为误差大小的可靠的后验误差估计, 用为停机准则.

注 2. 该定理可以直接推广到计算双线性形式的双边 Lanczos 方法 [12], 用双边 Lanczos 方法计算 $f(A)v$ 的近似, 称为 Lanczos 近似. 和 Arnoldi 过程相比, 双边的 Lanczos 过程虽然用的是短递推式, 但缺陷是数值稳定性差, 并可能出现恶性中断, 此时将不能计算 Lanczos 近似. 因此, 本文的讨论和实验仅限于 Arnoldi 近似.

4. $E_k(e^{-ht})$ 的一个上界估计

指数矩阵函数的双线性形式是应用中最常见的, 对于矩阵函数 $f(A)=e^{-hA}$, 取点列 $\{t_i\}_{i=0}^{\infty}$ 为零序列, 则指数函数的双线性形式 $u^Te^{-hA}v$ 的 Arnoldi 近似的误差展开式如下：

$$E_k(e^{-hA})=u^Te^{-hA}v-\beta u^TV_ke^{-hH_k}e_1=-h\beta h_{k+1,k}\sum_{j=1}^{\infty}e_k^T\phi_j(-hH_k)e_1u^T(-hA)^{j-1}v_{k+1}. \tag{4.1}$$

本节通过建立一个关于 $|E_k(e^{-hA})|$ 的精确上界估计, 首次定量给出误差展开式 (4.1) 的第一项是 $|E_k(e^{-hA})|$ 较好后验误差估计的理论依据.

由 (4.1) 可知其双线性形式 Arnoldi 近似误差向量如下：

$$E_k(e^{-hA})=u^Te^{-hA}v-\beta u^TV_ke^{-hH_k}e_1=-h\beta h_{k+1,k}\sum_{j=1}^{\infty}e_k^T\phi_j(-hH_k)e_1u^T(-hA)^{j-1}v_{k+1}. \tag{4.2}$$

记符号 $E_k^{(2)}(e^{-hA})$ 表示 $u^Te^{-hA}v$ 的误差展开式除第一项之外的剩余项之和, 即

$$E_k^{(2)}(e^{-hA})=-h\beta h_{k+1,k}\sum_{j=2}^{\infty}e_k^T\phi_j(-hH_k)e_1u^T(-hA)^{j-1}v_{k+1}. \tag{4.3}$$

则关于示 $u^Te^{-hA}v$ 的误差展开式的绝对值 $|E_k(e^{-hA})|$ 和剩余项之和的绝对值 $E_k^{(2)}(e^{-hA})$ 有下述定理成立.

定理 2. 设 A 为 n 阶实对称矩阵, 其谱 $\mathrm{spec}(A)$ 位于区间 $\Theta=[a,b]$ 中, $E_k^{(2)}(e^{-hA})$ 和 $E_k(e^{-hA})$ 分别有以下结果成立

$$|E_k^{(2)}(e^{-hA})|\le\gamma h\beta\eta_{k+1}|e_k^T\phi_1(-hT_k)e_1||u^Tv_{k+1}|, \tag{4.4}$$

$$|E_k(e^{-hA})|\le(1+\gamma)h\beta\eta_{k+1}|e_k^T\phi_1(-hT_k)e_1||u^Tv_{k+1}|, \tag{4.5}$$

其中 $\gamma=e^{h(b-a)}\gamma_1$, γ_1 是一个常数, 与矩阵 A 的谱 $\mathrm{spec}(A)$ 有关.

证明. 利用 [1] 定理 4.5 证明过程的中已有结果, 记关于矩阵函数乘向量 $f(A)v$ 的 Arnoldi 近似所得误差展开项除第一项之外的剩余项之和为 $e_k^{(2)}(e^{-hA})$. 已知

$$e_k^{(2)}(e^{-hA})=\beta\eta_{k+1}e_k^T\phi_1(-hT_k)e_1\left(\int_0^h\frac{z^k}{h^{k-1}}g(z)e^{(z-h)A}dz\right)Av_{k+1}, \tag{4.6}$$

其中 $|g(z)| \le e^{h(b-a)}$, 由此可得

$$\begin{aligned} |E_k^{(2)}(e^{-hA})| = |u^T e_k^{(2)}(e^{-hA})| &\le h\beta\eta_{k+1}|e_k^T\phi_1(-hT_k)e_1|e^{h(b-a)}|u^T\int_0^h e^{(z-h)A}dzAv_{k+1}| \\ &= h\beta\eta_{k+1}|e_k^T\phi_1(-hT_k)e_1|e^{h(b-a)}|u^T(I-e^{-hA})v_{k+1}| \\ &\le h\beta\eta_{k+1}|e_k^T\phi_1(-hT_k)e_1|e^{h(b-a)}(|u^Tv_{k+1}|+||e^{-hA}||) \\ &\le h\beta\eta_{k+1}|e_k^T\phi_1(-hT_k)e_1|e^{h(b-a)}\gamma_1|u^Tv_{k+1}| \\ &= \gamma h\beta\eta_{k+1}|e_k^T\phi_1(-hT_k)e_1||u^Tv_{k+1}|. \end{aligned} \tag{4.7}$$

因此对于 $E_k(e^{-hA})$ 有

$$\begin{aligned} |E_k(e^{-hA})| &\le h\beta\eta_{k+1}|e_k^T\phi_1(-hT_k)e_1||u^Tv_{k+1}| + |E_k^{(2)}(e^{-hA})| \\ &\le (1+\gamma)h\beta\eta_{k+1}|e_k^T\phi_1(-hT_k)e_1||u^Tv_{k+1}|, \end{aligned} \tag{4.8}$$

其中 $\gamma = e^{h(b-a)}\gamma_1$.

注 3. 对于 e^{-hA}, A 的特征值通常分布在右半平面. 因此, 在一般情况下, $|u^Tv_{k+1}| + ||e^{-hA}||$ 是与 $|u^Tv_{k+1}|$ 同量级的. 事实上, 当矩阵 A 的特征值远离原点时, 有 $||e^{-hA}|| \ll |u^Tv_{k+1}|$. 综上, 当矩阵的最小特征值不是特别靠近原点时, γ 中的因子 γ_1 是与 $O(1)$ 同量级的. h 在应用中远小于 1, 因此因子 $e^{h(b-a)} = O(1)$. 因此, 一般情况下, 误差展开式的第一项的绝对值可以作为指数函数双线性形式 $u^Te^{-hA}v$ 的 Arnoldi 近似的可靠的后验误差估计.

5. 数值实验

本节将用数值实验验证误差展开式的首项 $\xi_1 = \beta h_{k+1,k}|e_k^T\phi_1(H_k)e_1||u^Tv_{k+1}|$ 是真实误差一个较好的估计.

本文中所有的数值实验均在内存 4.0GB, 处理器 Intel(R) Core(TM) i7-6500U CPU @ 2.50 GHz(4 CPUs), 2.6GHz 的微机上实现, 系统是 Window 10, 使用软件为 MATLAB R2013a. 在计算 $\phi_1(H_k)$ 的时候需要用到 t_0 的值, 注意到 $t_0 \in S$ 且矩阵 H_k 的数值域属于 S, 不妨设 t_0 属于矩阵 H_k 的数值域, 实验过程中发现, ξ_1 的大小对不同 t_0 的选取并不敏感, 可取为 $h_{1,1}$ 或者是 0. 实验过程中需要计算 $f(H_k)$ 和 $\phi_1(H_k)$, 由于 H_k 是中小规模的, 因此我们直接采取 MATLAB 内置函数 funm 进行计算, 特别的当矩阵是实对称矩阵时, 采取谱分解的形式进行计算. 我们需要其与真实的误差进行比较, 实验过程中我们计算的是相对误差, 因此真实的相对误差为

$$\xi_{\text{true}}^{\text{rel}} = \frac{|u^Tf(A)v - \beta u^TV_kf(H_k)e_1|}{|u^Tf(A)v|}. \tag{5.1}$$

用定理 1 中的展开式的首项作为误差估计的相对误差估计 ξ^{rel} 为

$$\begin{aligned} \xi^{rel} &= \frac{\beta h_{k+1,k}|e_k^T\phi_1(H_k)e_1||u^Tv_{k+1}|}{|F_k|} \\ &= \frac{\beta h_{k+1,k}|e_k^T\phi_1(H_k)e_1||u^Tv_{k+1}|}{|\beta u^TV_kf(H_k)e_1|}. \end{aligned} \tag{5.2}$$

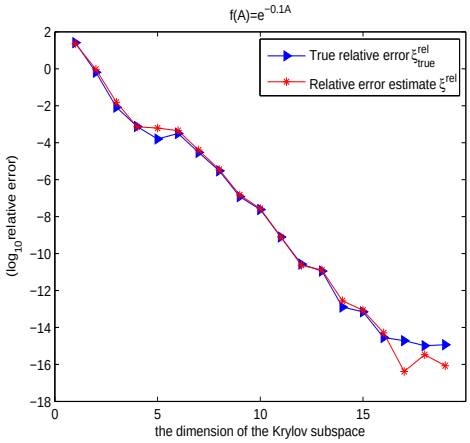
在 (5.2) 式的分母上, 由于真值 $u^Tf(A)v$ 未知, 我们采用计算得到的近似值 F_k 的绝对值 $|F_k|$ 来近似代替真实值 $|u^Tf(A)v|$, 这并不影响我们的实验目的, 因为在算法快收敛时, 两者的

值很接近. 下面我们将给出三种函数, 分别是指数函数 e^{-ht} 和双曲线 $1/t$ 以及三角函数的正、余弦函数 $\sin(-ht)$ 和 $\cos(-ht)$. 实验中取 $u \neq v$, 并分别用对称矩阵和非对称矩阵进行实验.

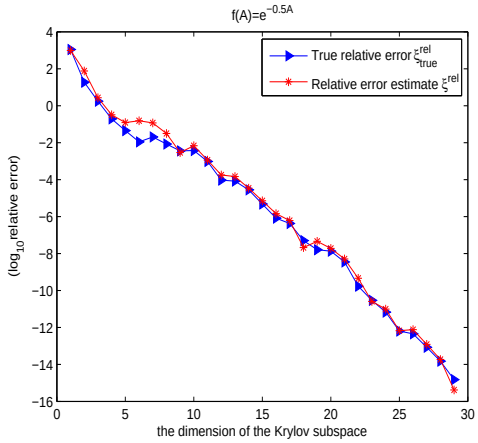
5.1. 矩阵函数 $f(A) = e^{-hA}$

本小节对指数函数的双线性形式进行实验, 分别针对对称矩阵与非对称矩阵.

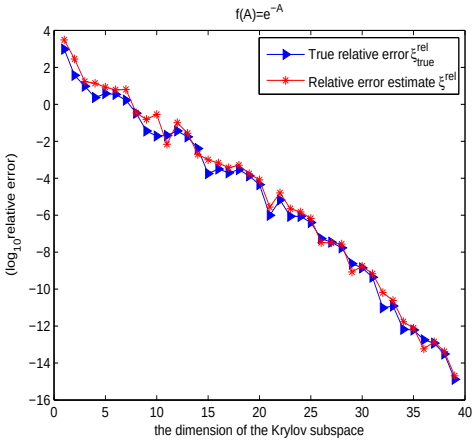
例 1. 此算例取自文献 [1]. 矩阵 A 为阶数 $n = 1001$ 的对角矩阵, 其特征值是均匀分布在 $[0, 40]$ 上, n 维列向量 u, v 分别由 MATLAB 随机生成, 并进行了单位化使得 $\|u\| = \|v\| = 1$, 分别取 $h = 0.1, 0.5, 1$, 比较 $u^T e^{-hA} v$ 的后验相对误差估计 ξ^{rel} 和真实相对误差估计 $\xi^{\mathrm{rel}}_{\mathrm{true}}$.



(a) 误差曲线 (h=0.1)



(b) 误差曲线 (h=0.5)



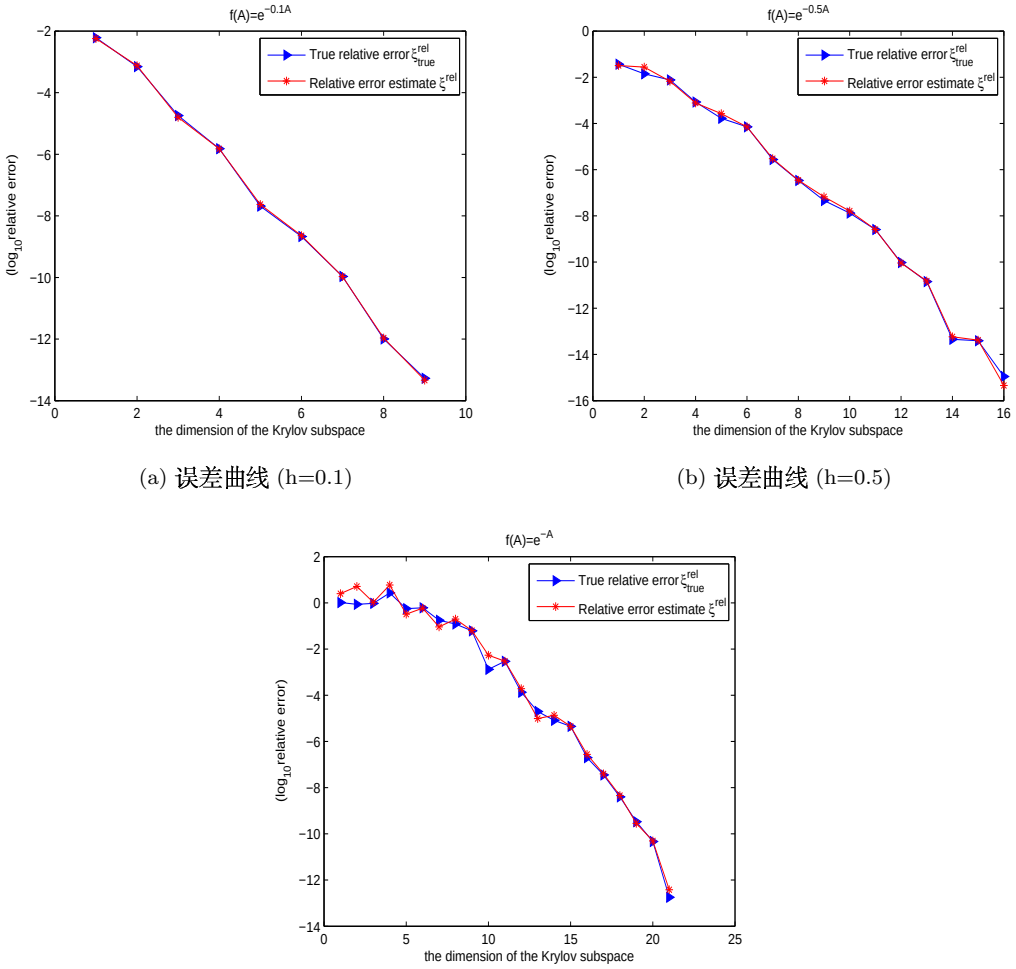
(c) 误差曲线 (h=1)

图 1: 数值计算对称矩阵 $u^T e^{-hA} v$ 的真实相对误差和后验相对误差比较

从图 1 中可以看出, 相对误差估计 ξ^{rel} 可以作为真实相对误差 $\xi^{\mathrm{rel}}_{\mathrm{true}}$ 的一个可靠的模拟, 并且发现 ξ^{rel} 与 $\xi^{\mathrm{rel}}_{\mathrm{true}}$ 贴合的非常好. 从实验结果中也看出 ξ^{rel} 并不受 γ 的大小影响. 此外,

发现当 h 越小时, 算法收敛的速度越快, 可见其收敛性与矩阵特征值的分布范围有关.

例 2. 该算例中的矩阵来源于 Higham 的测试集, 矩阵 A 为其中的矩阵 grcar, 阶数 $n = 2000$, n 维列向量 u, v 分别由 MATLAB 随机生成, 并进行了单位化使得 $\|u\| = \|v\| = 1$, 分别取 $h = 0.1, 0.5, 1$, 比较 $u^T e^{-hA} v$ 的 Arnoldi 后验相对误差估计 ξ^{rel} 和真实相对误差估计 ξ^{rel}_{true}.



(a) 误差曲线 (h=0.1)　　(b) 误差曲线 (h=0.5)

(c) 误差曲线 (h=1)

图 2: 数值计算非对称矩阵 $u^T e^{-hA} v$ 的真实相对误差和后验相对误差比较

由图 2 可知, 与算例 1 的结果类似, 当矩阵为非对称矩阵时, 后验相对误差 ξ^{rel} 与真实相对误差 $\xi^{\mathrm{rel}}_{\mathrm{true}}$ 的图像吻合的很好, 数值形态几乎完全一致, 因此对非对称矩阵来说, 误差展开式第一项的大小可以作为一个可靠的后验误差估计, 用作停机准则.

综上, 对于指数函数, 其双线性形式针对对称矩阵和非对称矩阵我们均能得到误差展开式的首项是一个可靠的后验误差估计, 并且设计为算法的停机准则. 这也是我们定理 2 所理论

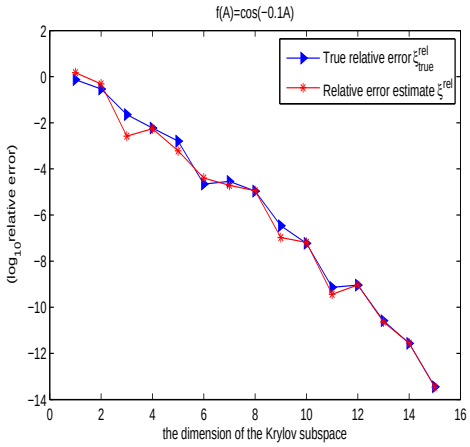
认证的.

对于其他解析函数, 譬如三角函数, 我们也有同样的结论.

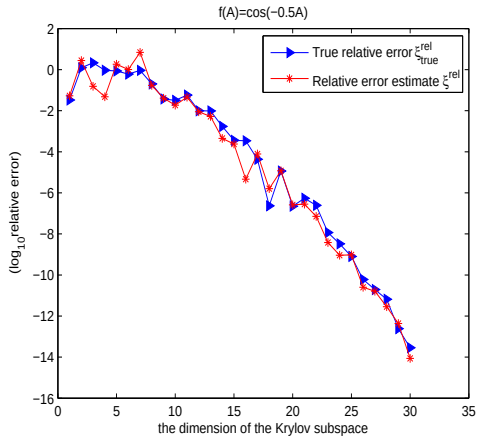
5.2. 矩阵函数 $f(A)=\cos(-hA)$

由定理 1 的证明过程中发现, 只要函数足够光滑, 均可以得到误差展开项的第一项大小可以作为算法可靠的后验误差估计. 因此这里我们考虑一类三角函数, 不失一般性, 我们选取余弦函数进行数值实验, 下面的数值实验将显示 ξ^{rel} 作为后验相对误差估计的可靠性. 这里依然针对对称矩阵和非对称矩阵进行实验.

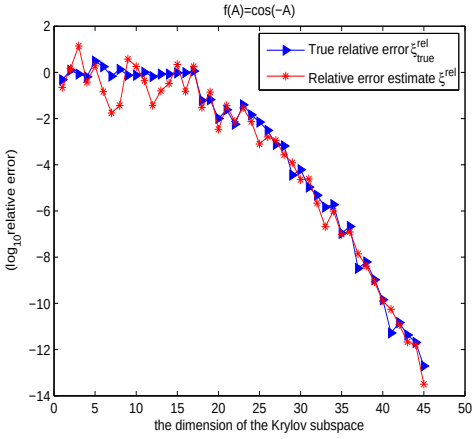
例 3. 选择算例 1 的对称矩阵, n 维列向量 u, v 分别由 MATLAB 随机生成, 并进行了单位化使得 $\|u\|=\|v\|=1$, 分别取 $h=0.1,0.5,1$, 比较 $u^T\cos(-hA)v$ 的后验相对误差估计 ξ^{rel} 和真实相对误差估计 $\xi^{\mathrm{rel}}_{\mathrm{true}}$, 结果如图 3 所示.



(a) 误差曲线 (h=0.1)



(b) 误差曲线 (h=0.5)



(c) 误差曲线 (h=1)

图 3: 数值计算对称矩阵 $u^T\cos(-hA)v$ 的真实相对误差和后验相对误差比较

对于余弦函数来说, 由图 3 发现, 当矩阵为对称矩阵时, 实验结果与指数函数的类似, 在算法刚开始的几步, $\xi^{\rm rel}$ 可能会高估或者低估真实相对误差 $\xi^{\rm rel}_{\rm true}$, 但并没有什么影响, 因为此时真实相对误差的精度也为 $O(1)$ 大小, 关键的是在算法开始收敛时, $\xi^{\rm rel}$ 与 $\xi^{\rm rel}_{\rm true}$ 的走势几乎一致了, 因此对于余弦函数来说, $\xi^{\rm rel}$ 依然可以作为矩阵函数双线性形式 $u^Tf(A)v$ 的 Arnoldi 算法的一个可靠的后验误差估计.

例 4. 该算例来源于文献 [2] 的 Example 2, 非对称矩阵 A 有复特征值且其分布在右半平面, 由于其特征值分布特征, 该算例经常被采用. 考虑初边值问题

$$\begin{aligned} \dot u - \Delta u + \delta_1 u_{x_1} + \delta_2 u_{x_2} = 0 \qquad & (x,t) \in (0,\,1)^3 \times (0,\,T), \\ u(x,t) = 0 \qquad & x \in \partial(0,\,1)^3,\ t \in [0,\,T], \\ u(x,0) = u_0(x), \qquad & x \in (0,\,1)^3. \end{aligned}$$

空间步长为 $\frac{1}{11}$, 离散化后得到如下形式的 ODEs:

$$\dot u(t) = -Au(t),\ t \in (0,\,T),\ u(0) = u_0.$$

阶数为 $n = 1000$ 非对称矩阵 A 可由 Kronecker 积表示：

$$A = -\frac{1}{11^2}[I_{10} \otimes (I_{10} \otimes F_1) + (E \otimes I_{10} + I_{10} \otimes F_2) \otimes I_{10}], \tag{5.3}$$

其中

$$E = \text{tridiag}(1, -2, 1),\ F_i = \text{tridiag}(1+\omega_i, -2, 1-\omega_i),\ \omega_1 = 3.2,\ \omega_2 \approx 4.27. \tag{5.4}$$

n 维列向量 u, v 分别由 MATLAB 随机生成, 并进行了单位化使得 $\|u\| = \|v\| = 1$, 取 $h = 1/121$, 比较 $u^T\cos(-hA)v$ 的后验相对误差估计 $\xi^{\rm rel}$ 和真实相对误差估计 $\xi^{\rm rel}_{\rm true}$, 误差曲线如图 4.

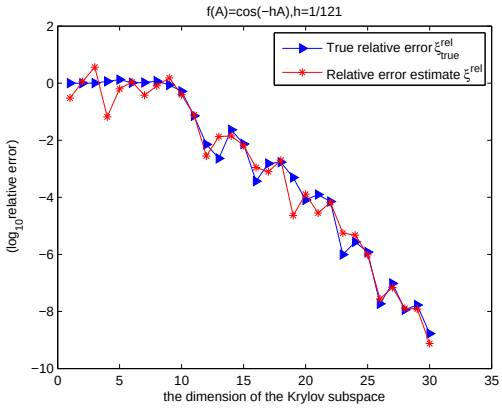


图 4: 数值计算非对称矩阵 $u^T\cos(-hA)v$ 的真实相对误差和后验相对误差比较

从图 4 中可以看出, 当矩阵为非对称矩阵时, 其结论和对称矩阵时相同, 因此, 无论矩阵 A 为对称矩阵还是非对称矩阵, $\xi^{\rm rel}$ 均可以作为余弦函数双线性形式 $u^Tf(A)v$ 的 Arnoldi 算法的一个可靠的相对后验误差估计.

5.3. 矩阵函数 $f(A) = A^{-1}$

当函数为双曲线 $1/t$ 时, 对应的矩阵函数的双线性形式为 $u^T A^{-1} v$. 该问题的计算有着十分广泛的应用背景, 譬如在模型降阶中, 更具体一点的来说, 在大规模集成电路设计系统中针对单输入单输出系统的传递函数就涉及到该双线性形式的计算. 关于该问题, 虽然我们不能在理论上证明误差展开式的首项是一个可靠的后验误差估计, 但是数值实验表明, 误差展开式的首项仍能很可靠地模拟误差的大小. 本小节的数值实验旨在说明我们得到的计算矩阵函数双线性形式 Arnoldi 算法的后验误差估计 ξ_1 具有实际的应用前景.

例 5. 该算例取自文献 [3] 中的 Example 5.1, 其中矩阵 A 是 MATLAB 中 *gallery* 中的 *parter* 矩阵, 阶数 $n = 1000$, n 维列向量 u, v 分别由 MATLAB 随机生成, 并进行了单位化使得 $\|u\| = \|v\| = 1$, 比较 $u^T A^{-1} v$ 的 Arnoldi 近似的后验相对误差估计 ξ^{rel} 和真实相对误差估计 $\xi^{\text{rel}}_{\text{true}}$, 误差曲线如图 5.

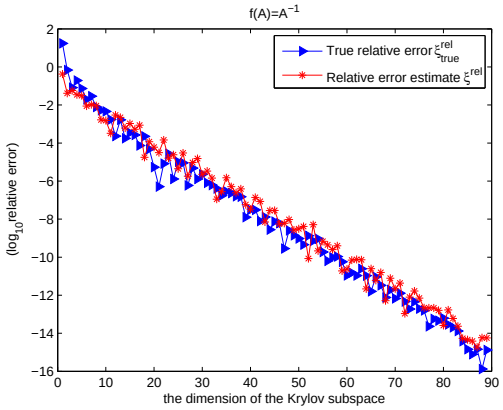


图 5: 数值计算非对称矩阵 $u^T A^{-1} v$ 的真实相对误差和后验相对误差比较

从图 5 中可以发现, 对于双线性形式 $u^T A^{-1} v$, ξ_1 作为误差展开式的第一项能可靠的模拟算法的收敛形态, 因此可以作为算法停机准则的一个可靠的相对后验误差估计.

综上, 对于满足定理 1 条件的足够光滑的函数来讲, 数值实验表明, 不管对称矩阵还是非对称矩阵, 误差展开式的第一项的大小均可以作为 Arnoldi 近似的一个可靠的后验误差估计, 从而可以作为算法的停机准则.

6. 总 结

本文旨在研究用 Arnoldi 方法矩阵函数的双线性形式并设计算法相应的停机准则. 具体来说, 就是将 $u^T f(A) v$ 看成是两个向量做内积, 其中一个向量 $f(A)v$ 是我们计算的关键. 而关于矩阵函数乘向量, 在 2015 年, Jia 和 Lv 已经给出了其 Krylov 子空间方法的可靠的后验误差估计. 本文将该结果推广到矩阵函数的双线性形式 $u^T f(A) v$ 的 Krylov 子空间方法的计算上. 我们给出了误差的误差展开式, 并证明其展开式的第一项可以作为可靠的后验误差估计. 具体地, 本文的工作内容主要有以下两点:

1. 介绍了计算矩阵函数的双线性形式 $u^T f(A)v$ 的 Krylov 子空间方法, 建立了算法误差的误差展开式, 给出了误差展开式的第一项大小可以作为后验误差估计的理论上的论证. 数值实验表明该误差展开式的第一项的大小可以作为可靠的后验误差估计, 从而可以在实际计算时设计为相应算法的停机准则.

2. 误差展开式的第一项可以作为算法的停机准则, 不仅仅针对于指数函数的双线性形式成立, 对于足够光滑的一类函数也有同样结论, 这些从数值实验上可以直观看出.

参 考 文 献

[1] 吕慧. 计算大规模稀疏矩阵函数乘向量的 Krylov 子空间算法 [D]. 清华大学, 2014.

[2] Jia Z, Lv H. A posteriori error estimates of Krylov subspace approximations to matrix functions. Numerical Algorithms, 2015, 69(1): 1–28.

[3] Fika P, Mitrouli M, Roupa P. Estimates for the bilinear form $x^T A^{-1} y$ with applications to linear algebra problems. Electronic Transactions on Numerical Analysis, 2014, 43: 70–89.

[4] Abramowitz M, Stegun I A. Handbook of mathematical functions: with formulas, graphs, and mathematical tables, volume 55. Courier Corporation, 1964.

[5] Bai Z, Fahey G, Golub G. Some large-scale matrix computation problems. Journal of Computational and Applied Mathematics, 1996, 74(1–2): 71–89.

[6] Bai Z, Golub G. Computation of large-scale quadratic forms and transfer functions using the theory of moments, quadrature and Padé approximation. Proceedings of Modern Methods in Scientific Computing and Applications. Springer, 2002: 1–30.

[7] Calvetti D, Kim S M, Reichel L. Quadrature rules based on the Arnoldi process. SIAM Journal on Matrix Analysis and Applications, 2005, 26(3): 765–781.

[8] Golub G. Matrix Computation and the Theory of Moments. Birkhäuser, Basel, 1995: 1440–1448.

[9] Golub G H, Meurant G. Matrices, moments and quadrature. Pitman Research Notes In Mathematics Series, 1994. 105–105.

[10] Golub G H, Meurant G. Matrices, moments and quadrature II; how to compute the norm of the error in iterative methods. BIT Numerical Mathematics, 1997, 37(3): 687–705.

[11] Golub G H, Strakoš Z. Estimates in quadratic formulas. Numerical Algorithms, 1994, 8(2): 241–268.

[12] Guo H, Renaut R A. Estimation of $u^T f(A)v$ for large scale unsymmetric matrices. Numerical Linear Algebra with Applications, 2004, 11(1): 75–89.

[13] 郭洪斌. 矩阵函数双线性形式的计算及其应用 [D]. 复旦大学, 1999.

[14] Strakoš Z. Model reduction using the Vorobyev moment problem. Numerical Algorithms, 2009, 51(3): 363–379.

THE ERROR ANALYSIS OF THE KRYLOV SUBSPACE METHODS FOR COMPUTING THE BILINEAR FORM OF MATRIX FUNCTIONS

Jia Zhongxiao Sun Xiaolin

(*Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China*)

Abstract

The bilinear form $u^T f(A)v$ of matrix functions is of wide interest in many applications, where $u, v \in \mathbb{R}^n, A \in \mathbb{R}^{n\times n}$, $f(z)$ is a given analytic function. In recent years, the efficient and reliable numerical algorithms for the bilinear form has been a research focus. Although there are numerous stopping criteria, they lack solid theoretical supports, and the reliability is unknown. In this paper, we consider the posteriori error estimates for the errors of approximate solutions of the matrix functions $u^T f(A)v$. We derive an error expansion and prove that the first term of the error expansion can used as a reliable stopping criterion.

Keywords: bilinear form; Krylov subspace method; relative error analysis; stopping criterion

2010 Mathematics Subject Classification: 65F30, 65F10