A Functionally Fitted 3-stage ESDIRK Method

Kazufumi Ozawa
Akita Prefectural University
Honjo Akita 015-0055, Japan
ozawa@akita-pu.ac.jp

Abstract

A special class of Runge-Kutta (-Nyström) methods called functionally fitted Runge-Kutta (FRK) methods has recently been proposed by the author. This class of methods is based on the exact integration of a given set of functions, which is called the basis functions by the author, and as a result, the method is always exact, if the solution of the ODE is expressed by a linear combination of the basis functions. In this paper, a 3-stage explicit singly diagonally implicit Runge-Kutta (ESDIRK) method in this class is developed. It is shown that the global error of the proposed method behaves like $O(h^4)$, where h is the step-size, even for the cases that the method is not fitted completely to the equation. The method is extended to an embedded pair. Several numerical experiments show that the method integrates the ODE very accurately, for the case that a suitable set of the basis functions can be found, and produces reasonable accuracy even for general cases.

1 Introduction

Initial value problems of ODEs are very important tools in science and technology. When solving the problems, it is often the case, in practice, that a priori information on the solution and/or equation, such as the period of the solution or the dominant eigenvalue of the coefficient matrix, is available. For this case, if we could design a numerical method based on such information, then the method would be very accurate for this problem. For example, if the solution of the ODE is a sinusoidal function with a small perturbation, then the special method which is exact only for the trigonometric function with the frequency will be more accurate than general ODE methods.

Many special methods which are exact for trigonometric functions, exponential functions, or mixed-polynomials have been derived (see e.g. [2], [7], [10], [11], [13], [14]). As an extension of these method, Ozawa has recently developed the Runge-Kutta (-Nyström) method that is exact on the linear space of any given functions ([8], [9]). The method is also an extension of the collocation Runge-Kutta methods, which are the special cases that the functions are polynomials. The method proposed by Ozawa, like collocation methods, is fully implicit, so that its computational cost is extremely expensive compared with explicit methods, and expensive with diagonally implicit Runge-Kutta (DIRK) methods.

The purpose of this work is to develop a computationally cheap Runge-Kutta method which are exact for the given functions, using the similar technique used in Ozawa ([8] and [9]).

2 Functionally fitted Runge-Kutta method

Consider the initial value problem

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = f(y(t)), \quad y(0) = y_0, \quad t \in [0,\, T], \tag{1}$$

and the s-stage Runge-Kutta method

$$\begin{cases} y_{n+1} = y_n + h \sum_{i=1}^{s} b_i \, f(Y_i), \\ \\ Y_i = y_n + h \sum_{j=1}^{s} a_{i,j} \, f(Y_j), \quad i = 1, \ldots, s, \end{cases} \tag{2}$$

where h is a step-size, and y_n is the numerical approximation to $y(nh)$.

Almost all Runge-Kutta methods are designed to integrate the equation exactly, if the solution of (1) is any polynomial of some degree or less. In our approach, however, the Runge-Kutta method is designed to be exact not for polynomials but for the linear space of given functions $\{\Phi_m(t)\}_{m=1}^{s}$. We call the functions $\{\Phi_m(t)\}_{m=1}^{s}$ the *basis functions*, and the resulting Runge-Kutta method and the *functionally fitted Runge-Kutta* (FRK) method.

One way to get the coefficients of the FRK is to give the functions $\{\Phi_m(t)\}_{m=1}^{s}$, and then solve the $(s+1)$ simultaneous equations

$$\begin{cases} \Phi_m(t + c_i \, h) = \Phi_m(t) + h \sum_{j=1}^{s} a_{i,j} \, \varphi_m(t + c_j \, h), & m \in \mathcal{F}_i, \quad i = 1, \ldots, s, \\ \\ \Phi_m(t + h) = \Phi_m(t) + h \sum_{i=1}^{s} b_i \, \varphi_m(t + c_i \, h), & m \in \mathcal{F}_{s+1}, \end{cases} \tag{3}$$

for the unknowns $a_{i,j}$ and b_i, where $\varphi_m(t) = \Phi_m'(t)$, and $\mathcal{F}_i \subseteq \mathcal{F} \equiv \{1, 2, \ldots, s\}$ is the set of the subscripts m of the functions $\Phi_m(t)$ used in the ith equation. Here we assume that c_i's are constant and different from each other. In [8] and [9], $\mathcal{F}_i$ is always $\mathcal{F}$ for each i, that is, all the functions $\Phi_m(t)\,(m = 1, \ldots, s)$ are used to determine all the unknowns $a_{i,j}\,(j = 1, \ldots, s)$, which means the resulting method is necessarily a fully implicit one. In the present case, however, we first give $s - |\mathcal{F}_i|$ coefficients in accordance with the sparsity pattern of a predetermined Butcher array, and then determine the remaining $|\mathcal{F}_i|$ coefficients using $|\mathcal{F}_i|$ different functions of $\Phi_1(t), \ldots, \Phi_s(t)$. It is clear from this construction that the method is always exact, if the solution $y(t)$ is an element of the space spanned by the $\Phi_\nu(t)$'s, where $\nu \in \cap_{i=1}^{s+1} \mathcal{F}_i$.

In Ozawa [8], the coefficients of FRK given by (3) are shown to be uniquely determined for all h and $t \in [0, T]$, if the Wronskian matrix

$$W(t) \equiv \begin{pmatrix} \varphi_1(t) & \cdots & \varphi_s(t) \\ \varphi_1^{(1)}(t) & \cdots & \varphi_s^{(1)}(t) \\ \vdots & \cdots & \vdots \\ \varphi_1^{(s-1)}(t) & \cdots & \varphi_s^{(s-1)}(t) \end{pmatrix} \tag{4}$$

is nonsingular. Moreover, in [8], these coefficients are shown to be analytic, if all of the functions $\{\Phi_m(t)\}_{m=1}^{s}$ are analytic in $[0, T]$.

In general, the coefficients $a_{i,j}$ and b_i determined in this way depend not only on h, but also on t. We shall consider, however, the case that these coefficients depend only on h; if the basis functions $\Phi_m(t)$ are polynomials, exponentials or sinusoidal functions, then this is the case. By this assumption, we set $t = 0$ in (3) without loss of generality.

3 Local truncation error of FRK method

The numerical results given by FRK will have truncation errors, except for the case that FRK is fitted to the problem (1) completely. Therefore, it is necessary to evaluate the error by some measure. In this paper, we use the "order of accuracy" to evaluate the error, as in the case of conventional numerical methods of ODEs. The definition of the order of accuracy of FRK is the same. That is, if the numerical solution by FRK satisfies

$$y_1 - y(h) = O(h^{p+1}), \quad y(0) = y_0, \quad h \to 0,$$

for any sufficiently smooth function $y(t)$, then we shall call the integer p the *order of accuracy* of FRK. However, unlike the conventional case of constant coefficient methods, we must consider the errors in the situation that the coefficients $a_{i,j}$ and b_i also vary as functions of h, when $h \to 0$.

In order to analyze the local truncation error of FRK, let us introduce the following quantities:

$$B(q) \equiv \sum_i b_i \, c_i^{q-1} - \frac{1}{q}, \tag{5}$$

$$C_i(q) \equiv \sum_j a_{i,j} \, c_j^{q-1} - \frac{c_i^q}{q}, \tag{6}$$

$$D(q) \equiv \sum_i b_i \, C_i(q). \tag{7}$$

According to Ozawa [8], [9], these quantities satisfy

$$\begin{aligned} B(q) &= O(h^{r_{s+1}+1-q}), \qquad q = 1, \ldots, r_{s+1}, \\ C_i(q) &= O(h^{r_i+1-q}), \qquad q = 1, \ldots, r_i, \end{aligned} \tag{8}$$

where we set $r_i = |\mathcal{F}_i| \, (i = 1, 2, \ldots, s+1)$.

We express the errors at the stages and step by using $B(q)$ and $C_i(q)$. First we consider the residuals at the stages and step. Let $y(t)$ be any sufficiently smooth function (not necessary the solution of (1)), then

$$\begin{aligned} R &\equiv y(0) + h \sum_i b_i \, y'(c_i h) - y(h) = \sum_{q \geq 1} \frac{h^q \, B(q)}{(q-1)!} \, (y'(0))^{(q-1)}, \\ R_i &\equiv y(0) + h \sum_j a_{i,j} \, y'(c_j h) - y(c_i h) = \sum_{q \geq 1} \frac{h^q \, C_i(q)}{(q-1)!} \, (y'(0))^{(q-1)}. \end{aligned} \tag{9}$$

Note that $y(t) = \Phi_m(t)$ these residuals vanish, which implies

$$\begin{aligned} \sum_{q \geq 1} \frac{h^q \, B(q)}{(q-1)!} \, (\varphi_m(0))^{(q-1)} &= 0, \quad m \in \mathcal{F}_{s+1} \\ \sum_{q \geq 1} \frac{h^q \, C_i(q)}{(q-1)!} \, (\varphi_m(0))^{(q-1)} &= 0, \quad m \in \mathcal{F}_i. \end{aligned} \tag{10}$$

On the other hand, if $\Phi_m(t)$ are polynomials of some degree or less, then $B(q)$ and $C_i(q)$ vanish for the first several q's, and $\varphi_m^{(q-1)}(t) = 0$ for the other q's. From (8) and (9) we have

$$R = O(h^{r+1}), \qquad R_i = O(h^{\rho+1}), \tag{11}$$

where

$$\rho = \min_i\{r_i\}, \quad r = r_{s+1}.$$

Next we consider the relation between the residuals and local errors of FRK method.

Let $y(t)$ be the solution of $y'(t) = f(y(t))$, then the errors at the stages are given by

$$\begin{aligned} e_i &\equiv Y_i - y(c_i\,h) \\ &= y_0 + h\sum_j a_{i,j}\,f(Y_j) - \left(y_0 + h\,\sum_j a_{i,j}\,y'(c_j\,h) - R_i\right) \\ &= h\,f_y\,\sum_j a_{i,j}\,(e_j + O(e_j^2)) + R_i, \end{aligned}$$

therefore

$$e_i = (1 - a_{i,\,i}\,h\,f_y)^{-1}\left((h\,f_y)\,\sum_{j\neq i} a_{i,j}\,(e_j + O(e_j^2)) + R_i\right) = O(h^{\rho+1}). \tag{12}$$

For the error at the step, we have

$$\begin{aligned} E &\equiv y_1 - y(h) \\ &= y_0 + h\sum_i b_i\,f(Y_i) - \left(y_0 + h\,\sum_i b_i\,y'(c_i\,h) - R\right) \\ &= h\sum_i b_i\,\left(f(Y_i) - f(y(c_i\,h))\right) + R \\ &= h\,f_y\,\sum_i b_i\,(Y_i - y(c_i\,h) + O(e_i^2)) + R. \end{aligned} \tag{13}$$

Before evaluating E, we must evaluate

$$\begin{aligned} \sum_i b_i\,Y_i &= \sum_i b_i\,y_0 + h\sum_{i,\,j} b_i\,a_{i,j}\,f(Y_j), \\ \sum_i b_i\,y(c_i\,h) &= \sum_i b_i\,y_0 + h\sum_{i,\,j} b_i\,a_{i,j}\,y'(c_j\,h) - T, \end{aligned}$$

where we put

$$T = \sum_i b_i\,R_i = \sum_{q\geq 1}\frac{h^q\,D(q)}{(q-1)!}\,(y'(0))^{(q-1)}. \tag{14}$$

For the order of T, if we assume

$$T = O(h^{\tau+1}), \tag{15}$$

then from (11) we have

$$\tau \geq \rho = \min_i\{r_i\}.$$

Thus

$$\begin{aligned} E &= (h\,f_y)\,\sum_{i,\,j} b_i\,a_{i,j}\,\left(f(Y_j) - y'(c_j\,h)\right) + (h\,f_y)\,T + R + O(h^{2\rho+3}) \\ &= (h\,f_y)^2\,\sum_{i,\,j} b_i\,a_{i,j}\,e_j + (h\,f_y)\,T + R + O(h^{2\rho+3}). \end{aligned}$$

If the order of $\sum_{i,j} b_i\, a_{i,j}\, e_j$ is that of the minimum of e_j's, then the order of accuracy p of the method is given by

$$p = \min\{\rho + 2,\, \tau + 1,\, r\}\,. \tag{16}$$

4 3-stage FESDIRK method

Let us consider the 3-stage Runge-Kutta method given by the Butcher array

$$\begin{array}{c|ccc} 0 & 0 & & \\ c_2 & a_{2,1} & \alpha & \\ c_3 & a_{3,1} & a_{3,2} & \alpha \\ \hline & b_1 & b_2 & b_3 \end{array} \tag{17}$$

Usually the methods of this type are called *explicit SDIRK* (ESDIRK) method when the coefficients are constant, and we shall call it *functionally fitted ESDIRK* (FESDIRK) method, if the method is FRK.

For the given c_i's and the given $\{\Phi_m(t)\}_{m=1}^3$, we determine the coefficients of (17) by

$$\begin{aligned} &\Phi_m(c_2\, h) = \Phi_m(0) + h\,(a_{2,1}\varphi_m(0) + \alpha\,\varphi_m(c_2\, h)), && m = 1, 2, \\ &\Phi_m(c_3\, h) = \Phi_m(0) + h\,(a_{3,1}\varphi_m(0) + a_{3,2}\,\varphi_m(c_2\, h) + \alpha\,\varphi_m(c_3\, h)), && m = 1, 2, \\ &\Phi_m(h) = \Phi_m(0) + h\,(b_1\varphi_m(0) + b_2\,\varphi_m(c_2\, h) + b_3\,\varphi_m(c_3\, h)), && m = 1, 2, 3, \end{aligned} \tag{18}$$

where we assume that the Wronskian matrix

$$W(t) = \begin{pmatrix} \varphi_1(t) & \varphi_2(t) & \varphi_3(t) \\ \varphi_1^{(1)}(t) & \varphi_2^{(1)}(t) & \varphi_3^{(1)}(t) \\ \varphi_1^{(2)}(t) & \varphi_2^{(2)}(t) & \varphi_3^{(2)}(t) \end{pmatrix} \tag{19}$$

is nonsingular at $t = 0$. This method is exact for $y(t) \in \mathrm{span}\{\Phi_1(t),\, \Phi_2(t)\}$. For this case, we have

$$r_2 = r_3 = 2, \quad r_4 = 3, \quad \rho = 2, \quad \tau \geq 2,$$

and

$$\begin{aligned} B(q) &= \sum_{i=1}^{3} b_i\, c_i^{q-1} - \frac{1}{q} = O(h^{4-q}), && q = 1, 2, 3, \\ C_i(q) &= \sum_{j=1}^{3} a_{i,j}\, c_j^{q-1} - \frac{c_i^q}{q} = O(h^{3-q}), && q = 1, 2, \end{aligned} \tag{20}$$

which leads to $p = 3$ from (16). We shall call the method FESDIRK3.

When $h \to 0$, FESDIRK approaches a constant coefficient method, which has a key role in later considerations. Here we consider the coefficients of this constant coefficient method. Relation (20) means that

$$\sum_{i=1}^{3} b_i^{(0)}\, c_i^{q-1} = \frac{1}{q}, \quad q = 1, 2, 3, \tag{21}$$

$$\sum_{j=1}^{i} a_{i,j}^{(0)}\, c_j^{q-1} = \frac{c_i^q}{q}, \quad q = 1, 2, \tag{22}$$

where $a_{i,j}^{(0)}$ and $b_i^{(0)}$ are the constant terms in the power series expansion in h of the coefficients. The relations (21) and (22), which are the so-called simplifying assumptions [1], determine $a_{i,j}^{(0)}$ and $b_i^{(0)}$ uniquely as functions of c_2. The results are:

$$\begin{cases} a_{2,1}^{(0)} = \dfrac{c_2}{2}, \quad a_{2,2}^{(0)} = \dfrac{c_2}{2}\,(=\alpha), \\ a_{3,1}^{(0)} = -\dfrac{36\,c_2^4 - 120\,c_2^3 + 134\,c_2^2 - 60\,c_2 + 9}{8\,c_2\,(3\,c_2-2)^2}, \\ a_{3,2}^{(0)} = -\dfrac{24\,c_2^3 - 50\,c_2^2 + 36\,c_2 - 9}{8\,c_2\,(3\,c_2-2)^2}, \quad a_{3,3}^{(0)} = \alpha, \\ b_1^{(0)} = \dfrac{6\,c_2^2 - 6\,c_2 + 1}{6\,c_2\,(4\,c_2-3)}, \\ b_2^{(0)} = \dfrac{1}{6\,c_2\,(6\,c_2^2 - 8\,c_2 + 3)}, \\ b_3^{(0)} = \dfrac{2\,(3\,c_2-2)^2}{3\,(4\,c_2-3)\,(6\,c_2^2 - 8\,c_2 + 3)}. \end{cases}$$

Note that $a_{i,j}^{(0)}$ and $b_i^{(0)}$ given above are independent of the choice of $\Phi_m(t)$.

5 Fourth order FESDIRK method

We have obtained a 3-stage FESDIRK method, which we call FESDIRK3, and have shown that the method is of order 3. In order to raise the order of the method up to 4 we assume two conditions.

The first condition is

$$\int_0^1 t^{q-1} \cdot t\,(t-c_2)\,(t-c_3)\,\mathrm{d}t \quad \begin{cases} = 0, & q = 1, \\ \neq 0, & q \geq 2. \end{cases} \tag{23}$$

We will consider later the case that the integral equals to 0 also for $q \geq 2$. From this assumption we have

$$c_3 = \frac{4\,c_2 - 3}{2\,(3\,c_2 - 2)}. \tag{24}$$

By assuming (24), we have from [8]

$$B(q) = \sum_{i=1}^{3} b_i\,c_i^{q-1} - \frac{1}{q} = O(h^{\max\{5-q,\,2\}}), \quad q = 1, \ldots, 4, \tag{25}$$

so that $r = 4$ in (11), and we have, instead of (21),

$$\sum_{i=1}^{3} b_i^{(0)}\,c_i^{q-1} = \frac{1}{q}, \quad q = 1, \ldots, 4. \tag{26}$$

The second assumption is

$$\sum_i b_i^{(0)}\,a_{i,\,j}^{(0)} = b_j^{(0)}\,(1 - c_j), \quad j = 1, 2, 3. \tag{27}$$

It has been shown that this condition together with (22) and (26) is a sufficient condition for the method $(a_{i,j}^{(0)}, b_i^{(0)}, c_i)$ to be of order 4 (see [1], [4]).

Next lemma proves that conditions (22), (26) and (27) guarantee $\tau = 3$ in (15).

LEMMA 1 *If conditions* (22), (26) *and* (27) *hold, then*

$$D(q) = O(h^{4-q}), \quad q = 1, 2, 3,$$

so that $\tau = 3$ *in* (15).

Proof. Let the power series expansion of $D(q)$ be

$$D(q) = D^{(0)}(q) + D^{(1)}(q)\, h + D^{(2)}(q)\, h^2 + \cdots$$

From the definition of $D(q)$ in (7) and the property of $C_i(q)$ given by (20), we have immediately

$$D(1) = O(h^2), \quad D(2) = O(h),$$

or equivalently

$$\begin{aligned} D^{(0)}(1) &= D^{(1)}(1) = 0, \\ D^{(0)}(2) &= 0. \end{aligned}$$

Next we show that several terms other than the above vanish. From (22), (26) and (27), we have for $q = 1, 2, 3$

$$D^{(0)}(q) = \sum_i b_i^{(0)} \left(\sum_j a_{i,j}^{(0)}\, c_j^{q-1} - \frac{c_i^q}{q} \right) = \sum_j b_j^{(0)} (1 - c_j)\, c_j^{q-1} - \frac{1}{q\,(q+1)} = 0. \tag{28}$$

On the other hand, from (14) and (18)

$$\begin{aligned} \sum_{q \geq 1} \frac{h^q\, D(q)}{(q-1)!} \left(\varphi_m(0)\right)^{(q-1)} &= \sum_{\nu \geq 1} \left(\sum_{q=1}^{\nu} \frac{\left(\varphi_m(0)\right)^{(q-1)}}{(q-1)!}\, D^{(\nu-q)}(q) \right) h^\nu \\ &= 0, \qquad m = 1, 2. \end{aligned} \tag{29}$$

Therefore, the condition that the coefficient of h^3 in (29) is being 0 can be expressed by

$$\begin{aligned} \varphi_1^{(0)}\, D^{(2)}(1) + \varphi_1^{(1)}\, D^{(1)}(2) + \frac{1}{2}\, \varphi_1^{(2)}\, D^{(0)}(3) = 0, \\ \varphi_2^{(0)}\, D^{(2)}(1) + \varphi_2^{(1)}\, D^{(1)}(2) + \frac{1}{2}\, \varphi_2^{(2)}\, D^{(0)}(3) = 0. \end{aligned} \tag{30}$$

Since we have already had $D^{(0)}(3) = 0$ in (28), and the submatrix

$$\begin{pmatrix} \varphi_1 & \varphi_2 \\ \varphi_1^{(1)} & \varphi_2^{(1)} \end{pmatrix} \tag{31}$$

of Wronskian matrix (19) is nonsingular by assumption, then

$$D^{(2)}(1) = 0, \quad D^{(1)}(2) = 0.$$

Summarizing the results obtained so far, we have

$$
\begin{aligned}
&D^{(0)}(1) = D^{(1)}(1) = D^{(2)}(1) = 0,\\
&D^{(0)}(2) = D^{(1)}(2) = 0,\\
&D^{(0)}(3) = 0.
\end{aligned}
$$

It is clear from the discussion in this lemma that any other terms of $D^{(l)}(q)$ never vanish. Thus we have proved this lemma.

Since $r = 4$ has already been established, and $\tau = 3$ has been proved by the above lemma, it is clear from (16) that $p = 4$. Thus we have the following theorem:

THEOREM 1 *If the abscissae c_2 and c_3 satisfy the two conditions* (24) *and* (27)*, then FESDIRK with the coefficients given by* (3) *is of order 4.*

Hereafter we call the (F)ESDIRK of order 4 (F)ESDIRK4. Next we must obtain the values of c_2 for which condition (27) is valid. Let d_j be

$$
d_j = \sum_i b_i^{(0)}\, a_{i,j}^{(0)} - b_j^{(0)}\,(1 - c_j), \quad j = 1, 2, 3,
$$

then from (21) and (22) we have

$$
\begin{aligned}
\sum_j d_j\, c_j^{q-1} &= \sum_{i,j} b_i^{(0)}\, a_{i,j}^{(0)}\, c_j^{q-1} - \sum_j b_j^{(0)}\,(1 - c_j)\, c_j^{q-1}\\
&= \frac{1}{q}\sum_i b_i^{(0)}\, c_i^q - \frac{1}{q} + \frac{1}{q+1} = 0, \quad \text{for } q = 1, 2,
\end{aligned}
$$

that is

$$
\begin{aligned}
d_1 + \quad d_2 + \quad & d_3 = 0,\\
c_2\, d_2 + \quad & c_3\, d_3 = 0.
\end{aligned}
$$

This means that if we force one of d_i's to be 0, then the remainders become 0, provided that $0 < c_2 \neq c_3$. Thus we put, for example,

$$
d_1 = -\frac{(3\,c_2 - 1)\,(3\,c_2 - 2)\,(c_2 - 1)}{6\,c_2\,(4\,c_2 - 3)} = 0,
$$

which leads to

$$
c_2 = \frac{1}{3}, \quad \frac{2}{3}, \quad 1.
$$

Among these solutions, $c_2 = 2/3$ is not allowed because of (24), so that we consider the remaining two solutions.

Next we show the stability regions of the ESDIRK4's with $c_2 = 1/3$ and $c_2 = 1$, and compare these regions with that of the classical Runge-Kutta method (RK4).

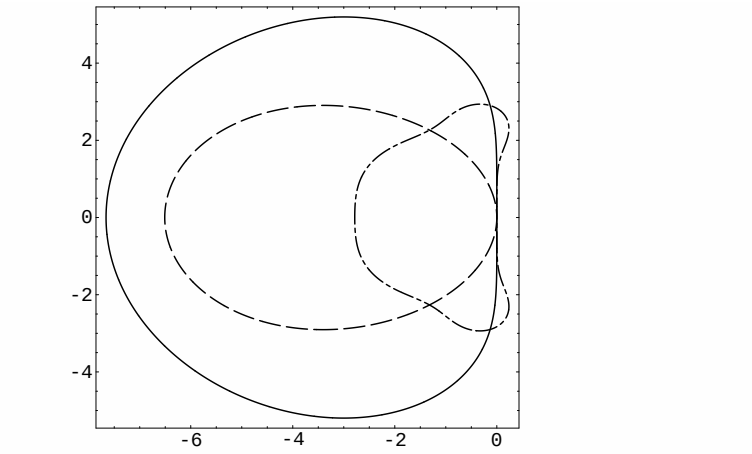

Fig. 1. Stability regions of ESDIRK4's with $c_2 = \frac{1}{3}$ (solid), $c_2 = 1$ (dashed), and RK4 (dash-and-dotted).

Fig. 1 shows that the ESDIRK4 with $c_2 = 1/3$ is preferable to the ESDIRK4 with $c_2 = 1$, since the former has broader stability region. Therefore we take $c_2 = 1/3$ also for FESDIRK4, since it is expected that FESDIRK has approximately the same properties as those of ESDIRK, when h is small. Here we show the Butcher array of the ESDIRK4 with $c_2 = 1/3$.

$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{6} & & \\ \frac{5}{6} & \frac{1}{24} & \frac{5}{8} & \frac{1}{6} & \\ \hline & \frac{1}{10} & \frac{1}{2} & \frac{2}{5} & \end{array} \tag{32}$$

Hereafter, we simply denote the ESDIRK4 and FESDIRK4 with $c_2 = 1/3$ by ESDIRK4 and FESDIRK4, respectively.

Finally we investigate the attainable order with the FESDIRK of the type (17). It is clear from the previous discussion that the FESDIRK and the ESDIRK have always the same order, since the latter corresponds to the particular case that $\Phi_1(t) = t, \Phi_2(t) = t^2, \Phi_3(t) = t^3$, and the discussion is independent of the choice of the basis functions. Therefore, we will consider the attainable order of the ESDIRK instead of that of the FESDIRK.

Any 3-stage method with $c_1 = 0$ cannot be of order 6, so that the attainable order of the ESDIRK of the type (17) will be at most 5. If the order of the method is 5, then the condition

$$\sum_i b_i^{(0)} c_i^{q-1} = \frac{1}{q}, \quad q = 1, \ldots, 5 \tag{33}$$

must be satisfied. The set of the abscissae satisfying the condition is given by

$$c_2 = \frac{6 - \sqrt{6}}{10}, \quad c_3 = \frac{6 + \sqrt{6}}{10},$$

which is obtained by solving

$$\int_0^1 t^q \, (t - c_2) \, (t - c_3) \, \mathrm{d}t = 0, \quad q = 1, 2.$$

Substituting the c_2 and c_3 obtained now into (33), and solving this for $b_i^{(0)}$, we have

$$b_1^{(0)} = \frac{1}{9}, \quad b_2^{(0)} = \frac{16+\sqrt{6}}{36}, \quad b_3^{(0)} = \frac{16-\sqrt{6}}{36},$$

and the values of $a_{i,\,j}^{(0)}$'s which satisfy (22) with these c_i's are given

$$a_{2,\,1}^{(0)} = \frac{6-\sqrt{6}}{20}, \quad a_{2,\,2}^{(0)} = \frac{6-\sqrt{6}}{20},$$
$$a_{3,\,1}^{(0)} = \frac{6+\sqrt{6}}{100}, \quad a_{3,\,2}^{(0)} = \frac{12+7\sqrt{6}}{50}, \quad a_{3,\,3}^{(0)} = \frac{6-\sqrt{6}}{20}.$$

Unfortunately, the set of the values listed above does not satisfy some of the order conditions even for $p = 4$. For example, the order condition for the tallest tree of order 4, which is given by

$$\sum_{i,j,k,l} b_i^{(0)}\, a_{i,j}^{(0)}\, a_{j,k}^{(0)}\, a_{k,l}^{(0)} = \frac{1}{24},$$

is not satisfied with these values; this value becomes $\frac{57-2\sqrt{6}}{1200}$ for the present set of the coefficients. Thus we have the conclusion that the attainable order with the FESDIRK is 4. This means that FESDIRK4 is the highest order method in the class of (17).

6 Numerical example

In order to see how well FESDIRK4 is fitted to the special problems for which we can find the basis functions, and whether or not the global error of the method behaves like $O(h^4)$ for general problems, we shall some present numerical examples. The problems to be solved are:

A. Airy equation

B. Bessel problem

C. Constant coefficient linear equation

D. Duffing equation

Problem A is the one whose solution oscillates with varying "frequency." Problems B and D are perturbed oscillations, and C is the problem whose solution consists of the two components: rapidly damped oscillatory component and decaying exponential component. In order to generate the coefficients of FESDIRK4, we use sinusoidal bases for problems A, B and D, and exponential bases for problem C. In these experiments, we compare the Euclidean norm of the errors with those of the other methods. All the computations are performed by the IEEE double precision arithmetic.

Airy equation

Consider the Airy equation

$$y''(t) - t\, y(t) = 0, \tag{34}$$

with the initial condition

$$\begin{aligned} y\,(-50) &= \mathrm{Ai}(-50) + 0.5\,\mathrm{Bi}(-50) &&= -2.304564997\cdots\times 10^{-1}, \\ y'(-50) &= \mathrm{Ai}'(-50) + 0.5\,\mathrm{Bi}'(-50) &&= \ \ 3.963089871\cdots\times 10^{-1}, \end{aligned}$$

where $\mathrm{Ai}\,(t)$ and $\mathrm{Bi}\,(t)$ are Airy's Ai and Bi functions, which are linearly independent solutions of Eq. (34) (see [6]). The exact solution of the problem is

$$y(t) = \mathrm{Ai}(t) + 0.5\,\mathrm{Bi}(t).$$

For this problem, the basis functions

$$\Phi_1(t) = t, \quad \Phi_2(t) = \cos(\omega\,t), \quad \Phi_3(t) = \sin(\omega\,t), \tag{35}$$

will be appropriate, in which case the Wronskian matrix associated with the functions is nonsingular. We integrate the equation from $t = -50$ to 0, changing the angular frequency ω by the formula

$$\omega = \sqrt{-t},$$

at every integer point $t = -50, -49, \ldots, 0$. The error of FESDIRK4 is compared favourably with that of ESDIRK4 in Fig. 1, although both of the methods are 4th order accurate.

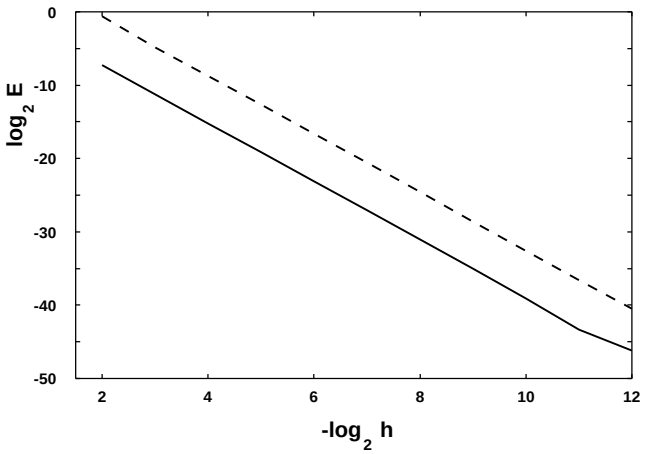

Fig. 1. Errors E of FESDIRK4 (solid) and ESDIRK4 (dashed) versus step-size h for Airy equation (34).

Bessel problem [15]

Next, let us consider the equation

$$y''(t) + 100\,y(t) = -\frac{y(t)}{4\,t^2}, \tag{36}$$

with the initial condition

$$\begin{aligned} y\,(0.5) &= \sqrt{0.5}\,J_0(5) &&= -1.255798813\cdots\times 10^{-1} \\ y'(0.5) &= \frac{J_0(5)}{2\sqrt{0.5}} - 10\,\sqrt{0.5}\,J_1(5) &&= \ \ 2.190754414\cdots \end{aligned}$$

The exact solution of the problem is given by

$$y\,(t) = \sqrt{t}\,J_0(10\,t),$$

where $J_\nu(t)$ and $Y_\nu(t)$ are the Bessel functions of the first and second kind, respectively. We integrate the equation from $t = 0.5$ to 10 by the two methods: FESDIRK4 with $\omega = 10$ (fixed) and ESDIRK4. The results are shown in Fig. 2.

From the result, we can observe that the two methods are of order 4 also in this example, and that the accuracy of FESDIRK4 is remarkable compared with the previous example.

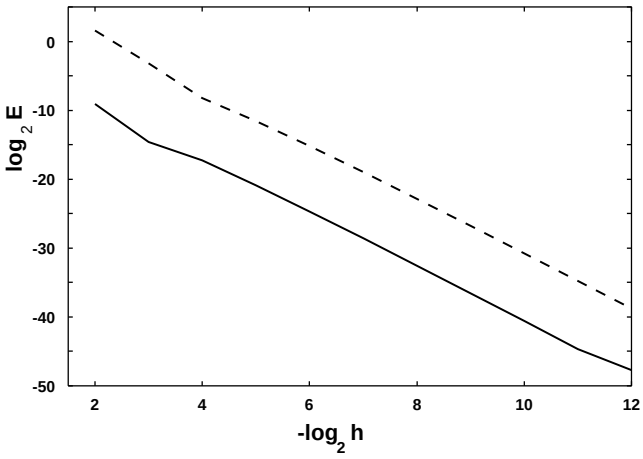

Fig. 2. Errors E of FESDIRK4 (solid) and ESDIRK4 (dashed) versus step-size h for Bessel problem (36).

Constant coefficient linear equation

The third problem to be considered is the linear homogeneous equation

$$y'(t) = P\,y(t), \quad y(0) = (1,\,0,\,0,\,0)^{\mathrm{T}}, \tag{37}$$

where

$$P = \begin{pmatrix} 0 & 0 & 1 & 101 \\ -96 & -1 & -97 & 6 \\ -98 & 0 & -99 & -96 \\ -1 & 0 & -1 & -102 \end{pmatrix}.$$

The exact solution of the problem is given by

$$y(t) = \begin{pmatrix} \mathrm{e}^{-t} + \mathrm{e}^{-100\,t} \sin t \\ \mathrm{e}^{-t}\,(-1+t) + \mathrm{e}^{-100\,t}\,(\cos t + 2\sin t) \\ -\mathrm{e}^{-t} + \mathrm{e}^{-100\,t}\,(\cos t + \sin t) \\ -\mathrm{e}^{-100\,t}\,\sin t \end{pmatrix}.$$

This solution consists of fast and slow modes. If the step-size in the stability region is used, then the fast mode will be damped out very soon and, as a result, the slow mode will dominate the entire solution. Hence, we fit the method to the slow mode, that is, we choose the following basis functions:

$$\Phi_1(t) = t, \quad \Phi_2(t) = \exp(-t), \quad \Phi_3(t) = t\,\exp(-t). \tag{38}$$

We integrate the equation from $t = 0$ to 2 by the FESDIRK4 with basis functions (38), and compare the error with those of the three Runge-Kutta methods: ESDIRK4, the

2-stage Gauss (Gauss2) and the classical Runge-Kutta (RK4) methods, each of which is of order 4. The results are shown in Table 1.

Table 1. Errors of various methods for linear equation (37).

$-\log_2 h$	$\log_2 E$			
	FESDIRK4	ESDIRK4	Gauss2	RK4
2	2.708e+01	2.915e+01	-5.124e+00	1.099e+02
3	2.486e+01	2.713e+01	-2.196e+01	1.531e+02
4	-2.858e+01	-2.585e+01	-2.529e+01	1.682e+02
5	-5.334e+01	-2.985e+01	-2.929e+01	4.702e+01
6	-5.271e+01	-3.387e+01	-3.329e+01	-3.068e+01
7	-5.262e+01	-3.787e+01	-3.729e+01	-3.470e+01
8	-5.125e+01	-4.188e+01	-4.129e+01	-3.870e+01
9	-5.091e+01	-4.586e+01	-4.530e+01	-4.270e+01
10	-5.164e+01	-5.073e+01	-4.907e+01	-4.668e+01
11	-5.212e+01	-5.056e+01	-5.232e+01	-5.163e+01
12	-5.016e+01	-5.062e+01	-4.988e+01	-5.078e+01

E is the Euclidean norm of the error at $t = 2$.

It can been seen that, although FESDIRK4 is less stable than the 2-stage Gauss Runge-Kutta method for large step-sizes, this method is fitted to the problem completely for moderately small step-sizes; the values of order -50 or less in the second column of the table must be the accumulations of round-off errors, since the machine epsilon of the IEEE double precision arithmetic is 2^{-53}. On the other hand, the errors of the other methods decrease very slowly at the rate of $O(h^4)$.

Duffing equation [3]

The last one to be integrated by FESDIRK4 is a nonlinear equation. Let us consider the Duffing equation

$$y''(t) + \mu^2\, y(t) = -k^2\,(y(t) - 2\, y(t)^3), \qquad (39)$$
$$y(0) = 0, \quad y'(0) = \mu.$$

The exact solution is given by

$$y(t) = \mathrm{sn}\,(\mu\, t; (k/\mu)^2),$$

where $\mathrm{sn}(\cdot\,;\cdot)$ is the Jacobian elliptic function. We integrate the equation with $\mu = 1$ and $k = 0.03$ from $t = 0$ to 100 by the FESDIRK4 with the basis functions (35)(we set $\omega = 1$) and ESDIRK4. The result is shown in Fig. 3.

For this example, both of the methods are more accurate compared with Problems A and B, in spite of the long term interval of integration.

To summarize, FESDIRK4 is a very efficient scheme for the special problems for which we can find the basis functions successfully, and is reasonably accurate for general problems, unless the problem is very stiff. This is due to the fact that the method has 4th order accuracy for general problems.

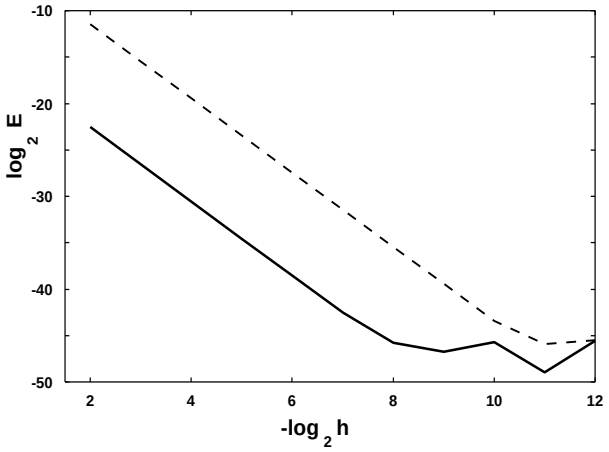

Fig. 3. Errors E of FESDIRK4 (solid) and ESDIRK4 (dashed) versus step-size h for Duffing equation (39).

7 Embedded FRK method

Since we have solved Problems A, B, C, and D, next we must consider 'E' (embedded) method. Let us consider the embedded pair

$$\begin{aligned} y_{n+1} &= y_n + h\,(b_1 f(Y_1) + b_2 f(Y_2) + b_3 f(Y_3)), \\ \bar{y}_{n+1} &= y_n + h\,(\bar{b}_1 f(Y_1) + \bar{b}_2 f(Y_2) + \bar{b}_3 f(Y_3) + \bar{b}_4\, f(Y_4)), \end{aligned} \tag{40}$$

where

$$\left\{ \begin{array}{ll} Y_1 & = y_n, \\ Y_2 & = y_n + h\,(a_{2,1}\, f(Y_1) + \alpha\, f(Y_2)), \\ Y_3 & = y_n + h\,(a_{3,1}\, f(Y_1) + a_{3,2}\, f(Y_2) + \alpha\, f(Y_3)), \\ Y_4 & = y_n + h\,(a_{4,1}\, f(Y_1) + a_{4,2}\, f(Y_2) + a_{4,3}\, f(Y_3) + \alpha\, f(Y_4)), \end{array} \right.$$

and we assume $c_2 = 1/3$ and $c_3 = 1$, as before. The Butcher array of the pair is

$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{3} & a_{2,1} & \alpha & & \\ \frac{5}{6} & a_{3,1} & a_{3,2} & \alpha & \\ 1 & a_{4,1} & a_{4,2} & a_{4,3} & \alpha \\ \hline & b_1 & b_2 & b_3 & 0 \\ \hline & \bar{b}_1 & \bar{b}_2 & \bar{b}_3 & \bar{b}_4 \end{array} \tag{41}$$

In the array, we further assume

$$\bar{b}_1 = a_{4,1}, \quad \bar{b}_2 = a_{4,2}, \quad \bar{b}_3 = a_{4,3}, \quad \bar{b}_4 = \alpha,$$

so that the method to calculate $\bar{y}_{n+1}$ becomes FSAL (first same as last). The computational cost of this method is approximately the same as that of FESDIRK4, since the number of LU decomposition to be performed in the step is still one.

Here we must determine the coefficients of the method. We take the same $a_{i,j}, \alpha$ and b_i as those of FESDIRK4, so that the order p of the method corresponds to b_i is 4. With

these coefficients, if we choose the $\bar{b}_i$ such that the order of the method corresponds to $\bar{b}_i$ is 3.

If we force

$$\bar{B}(q) \equiv \sum_{i=1}^{4} \bar{b}_i \, c_i^{q-1} - \frac{1}{q} = O(h^{4-q}), \quad q = 1, 2, 3, \tag{42}$$

then we have

$$\bar{R} \equiv y(0) + h \sum_{i=1}^{4} \bar{b}_i \, y'(c_i h) - y(h) = \sum_{q \geq 1} \frac{h^q \, \bar{B}(q)}{(q-1)!} \, (y'(0))^{(q-1)} = O(h^4).$$

Therefore, if we set $\bar{R} = O(h^{\bar{r}+1})$, then $\bar{r} = 3$ and

$$\bar{p} \equiv \min\{\rho + 2, \, \tau + 1, \, \bar{r}\} = \min\{4, \, 4, \, 3\} = 3. \tag{43}$$

The coefficients $\bar{b}_1$, $\bar{b}_2$ and $\bar{b}_3$ satisfying (42) are given by solving the system of the equations

$$\Phi_m(h) = \Phi_m(0) + h \, (\bar{b}_m \, \varphi_m(0) + \bar{b}_2 \, \varphi_m(c_2 \, h) + \bar{b}_3 \, \varphi_m(c_3 \, h) + \alpha \, \varphi_m(h)), \quad m = 1, 2, 3,$$

for the given $\{\Phi_m(t)\}_{m=1}^{3}$. Thus, we have the 3rd order method embedded in the 4th order one. The step-size strategy for this pair, which controls the local truncation error of the lower order method within a prescribed tolerance TOL, is given by

$$h_{n+1} = \theta \left(\frac{TOL}{\|\bar{y}_n - y_n\|} \right)^{1/4} h_n,$$

where θ is a safety factor, say $\theta = 0.9$.

Now, let us apply the embedded pair to the two-body problem [5], [12]

$$y_1'' = -y_1/r^3, \quad y_2'' = -y_2/r^3, \quad r = \sqrt{y_1^2 + y_2^2}, \tag{44}$$

with the initial condition

$$y_1(0) = 1 - e, \quad y_2(0) = 0, \quad y_1'(0) = 0, \quad y_2'(0) = \sqrt{\frac{1+e}{1-e}},$$

where $e \, (0 \leq e < 1)$ is an eccentricity. The exact solution of this problem is

$$y_1(t) = \cos u - e, \quad y_2(t) = \sqrt{1 - e^2} \, \sin u,$$

where u is the solution of Kepler's equation

$$u = t + e \sin u.$$

The solution of (44) is found to be 2π-periodic for any e. Hence, a natural choice of the basis functions is (35) with $\omega = 1$. By this choice, the problem with small e is expected to be accurately solved, since the solution is purely sinusoidal when $e = 0$. We integrate the problem with $e = 0.005$ from $t = 0$ to $50 \, \pi$ by the two embedded pairs FESDIRK4(3) and the ESDIRK4(3).

From the result of Table 2, we can see that the embedded method derived here controls the local truncation error well, and as a result, the method integrates the equation with fewer steps compared with ESDIRK4(3).

Table 2. Errors and the total steps for two-body problem (44) with $e = 0.005$.

	FESDIRK4(3)		ESDIRK4(3)	
TOL	error	steps	error	steps
10^{-2}	2.785e+00	225	2.483e+00	136
10^{-3}	2.866e-01	170	2.153e+00	277
10^{-4}	7.846e-03	225	1.494e-01	496
10^{-5}	1.399e-03	381	9.359e-03	884
10^{-6}	1.690e-04	680	6.200e-04	1573
10^{-7}	1.846e-05	1207	4.416e-05	2796
10^{-8}	1.938e-06	2144	3.412e-06	4970
10^{-9}	1.993e-07	3806	2.848e-07	8833
10^{-10}	2.021e-08	6762	2.530e-08	15706

TOL: Tolerance of the local error.
error: Euclidean norms of the errors at $t = 50\,\pi$.
steps: Total number of time steps (including rejected steps).

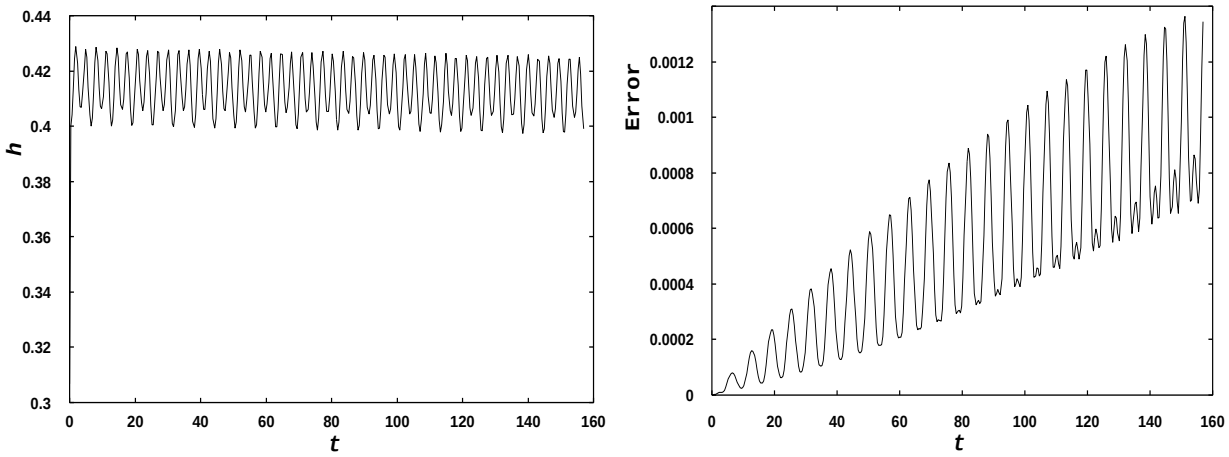

Fig. 5. Step-size plot and error behavior of embedded pair FESDIRK4(3).

8 Summary and future work

We have presented a functionally fitted 3-stage ESDIRK method. Although the method is of order 4 for general cases, the method is always exact when the solution of the ODE can be expressed by a linear combination of the given basis functions. Various numerical examples show that the method has proved successful when a suitable set of the basis functions is found, and reasonably accurate, even if this is not the case. The method is extended to an embedded pair.

The stability analysis and the implementation issue of FRK will be future works.

Acknowledgment. The author would greatly appreciate the conference organizers, Prof. S. Maeda, Prof. T. Ito and Prof. T. Mitsui, for their efforts. The author also wishes to thank Mr. C. Hirota, the research assistant of Akita Prefectural University. He provided much support for this work.

References

[1] J. Butcher, *The Numerical Analysis of Ordinary Differential Equations*, Wiley, 1987.

[2] J.P. Coleman, Mixed interpolation methods with arbitrary nodes, J. Comput. Appl. Math. **92** (1998), 69–83.

[3] J.M. Franco, Embedded pairs of explicit ARKN methods for the numerical integration of perturbed oscillators, Proceedings of the 2002 Conference on Computational and Mathematical Methods on Science and Engineering CMMSE-2002, (Sep. 2002, at Alicante Spain), Vol 1. 92–101.

[4] E. Hairer, S.P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I*, Springer-Verlag, Second Revised Edition, 1992.

[5] T.E. Hull, W.H. Enright, B.M. Fellen and A.E. Sedgwick, Comparing numerical methods for ordinary differential equations, SIAM J. Numer. Anal., **9** (1972), 603–637.

[6] N.N. Lebedev, *Special Functions & Their Applications* (Translated & edited by R.A. Silverman), Dover Publications, Inc. 1972.

[7] K. Ozawa, A four-stage implicit Runge-Kutta-Nyström method with variable coefficients for solving periodic initial value problems, Japan Journal of Industrial and Applied Mathematics, **16** (1999), 25–46.

[8] K. Ozawa, Functional fitting Runge-Kutta method with variable coefficients, Japan Journal of Industrial and Applied Mathematics **18** (2001), 105–128.

[9] K. Ozawa, Functional fitting Runge-Kutta-Nyström method with variable coefficients, Japan Journal of Industrial and Applied Mathematics **19** (2002), 55–85.

[10] K. Ozawa, Functionally fitted linear multistep method, Proceedings of the 2002 Conference on Computational and Mathematical Methods on Science and Engineering CMMSE-2002, (Sep. 2002, at Alicante Spain), Vol 1. 271–280.

[11] B. Paternoster, Runge-Kutta (-Nyström) methods for ODEs with periodic solutions based on trigonometric polynomials, Applied Numerical Mathematics **28** (1998), 401–412.

[12] F.L. Shampine, *Numerical Solution of Ordinary Differential Equations*, Chapman & Hall, 1994.

[13] T.E. Simos, A fourth algebraic order exponentially-fitted Runge-Kutta method for the numerical solution of the Schrödinger equation, IMA J. Numer. Anal. **21** (2001), 919–931.

[14] G. Vanden Berghe, H. De Meyer, M. Van Daele and T. Van Hecke, Exponentially-fitted Runge-Kutta methods, Journal of Computational and Applied Mathematics, **125** (2000), 107–115.

[15] P.J. Van Der Houwen and B.P. Sommeijer, Diagonally implicit Runge-Kutta-Nyström methods for oscillatory problems, SIAM J. Numer. Anal. **26** (1989), 414–429.