Monografías Matemáticas *García de Galdeano* **39**, 147–154 (2014)

PROJECTION METHODS BASED ON DISPERSION ERRORS FOR RK METHODS

María Pilar Laburta and Juan Ignacio Montijano

Abstract. In this article a projection technique based on the dispersion error for Runge–Kutta (RK) methods is presented. In particular, we study how to apply it to the Bogacki–Shampine method of order 3, giving an algorithm that computes appropriate directions to project that RK method preserving some first integral of the differential system. Some numerical experiments are also carried out to show the efficiency of the new projection method.

Keywords: Projection methods, Runge–Kutta methods, invariant preservation.

AMS classification: 65L05, 65L06.

§1. Introduction

Let us consider autonomous differential systems of the form

$$y'(t) = f(y(t)), \tag{1.1}$$

where $f\colon D \subseteq \mathbb{R}^N \to \mathbb{R}^N$, is a sufficiently smooth function.

First integrals of (1.1) play an important role in the qualitative and quantitative study of the flow of these systems. A scalar function $G\colon \widehat{D} \subseteq \mathbb{R}^N \to \mathbb{R}$ of class $C^{(1}(\widehat{D})$, $\widehat{D} \subseteq D$, is a first integral of (1.1) if $\nabla G(y) f(y) = 0\ \forall y \in \widehat{D}$ [7, pp. 93]. As a consequence, if $y(t)$ is a solution of (1.1), then $G(y(t))$ is a constant quantity for all t.

It is natural to look for numerical approximations reproducing some desirable properties of the true solution of the differential system (see e.g. [3], [4], [5]). In this work we will deal with one-step numerical methods that preserve first integrals of (1.1). There are several techniques to do that as it can be seen in the introduction of [3]. Following the ideas of this article by Calvo et al., in this paper we will consider directional projection methods. If $y(t)$ denotes the solution of (1.1) satisfying $y(t_0) = y_0$, those projection methods provide approximations y_n to $y(t_n)$, with $t_n = t_0 + nh$, and h the step size, given by

$$y_{n+1} = \widetilde{y}_{n+1} + \lambda_n w_n,\ n = 0, 1, 2, \ldots, \tag{1.2}$$

where $\widetilde{y}_{n+1}$ is the numerical approximation to $y(t_{n+1})$ provided by a given explicit Runge–Kutta (RK) method, $w_n \in \mathbb{R}^N$ defines the direction of the projection, and λ_n is an scalar chosen so that $y_{n+1} \in \{y \in \mathbb{R}^N \,|\, G(y) = G(y_0)\}$, which means that y_{n+1} is the projection of $\widetilde{y}_{n+1}$ onto that variety. Thus, if we denote $g(y) := G(y) - G(y_0)$, λ_n will be calculated at each step by solving the equation:

$$g(\widetilde{y}_{n+1} + \lambda_n w_n) = 0.$$

§2. Projected Bogacki–Shampine method

Throughout this paper, we will take as the basic RK formula $\widetilde{y}_{n+1}$ in (1.2), the 3-stage, 3rd-order method obtained by Bogacki and Shampine in [1]. As in [3], we will take

$$w_n = \widehat{y}_{n+1} - \widetilde{y}_{n+1}, \tag{2.1}$$

where $\widehat{y}_{n+1}$ will be a consistent explicit RK method, embedded to $\widetilde{y}_{n+1}$, that will be chosen at each step. Thus, we have the embedded RK pair:

$$\begin{array}{c|c} c & A \\ \hline & \widetilde{b}^T \\ \hline & \widehat{b}^T \end{array} \quad = \quad \begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 3/4 & 0 & 3/4 & \\ \hline & 2/9 & 1/3 & 4/9 \\ \hline & \widehat{b}_1 & \widehat{b}_2 & 1-\widehat{b}_1-\widehat{b}_2 \end{array} \tag{2.2}$$

In order to choose appropriately the coefficients $\widehat{b}_1, \widehat{b}_2$, we firstly consider the dispersion error function [8, 2]. For an s-stage RK method with coefficients (A, b), $A \in \mathbb{R}^{s\times s}$, $b \in \mathbb{R}^s$, it is defined after comparing the exact solution of the scalar test equation $y' = i\omega y$, with $\omega \in \mathbb{R}$, which satisfies $y(t_{n+1}) = e^{i\nu}y(t_n)$, with the numerical solution provided by the RK method with fixed step size h, which satisfies $y_{n+1} = R(i\nu)y_n$, being $\nu = h\omega$, and $R(z)$ the stability function. More specifically, the dispersion error is given by

$$\phi(\nu) := \nu - \arg(R(i\nu)) = \nu - \arctan\frac{\mathrm{Im}(R(i\nu))}{\mathrm{Re}(R(i\nu))},$$

with $R(i\nu) = 1 + i\nu b^T(I - i\nu A)^{-1}e$, $i = \sqrt{-1}$, I the identity matrix of order s and $e = (1, \dots, 1)^T \in \mathbb{R}^s$. Furthermore, if $\phi(\nu) = O(\nu^{q+1})$, $\nu \to 0$, then the RK method is said to be dispersive of order q.

A desired property for $\widehat{y}_{n+1}$ is that if it advances the phase with respect to $y(t_{n+1})$, then $\widetilde{y}_{n+1}$ must delay it, and conversely, i.e.

$$\widehat{\phi}(\nu)\widetilde{\phi}(\nu) < 0, \quad (\nu \to 0). \tag{2.3}$$

Moreover, since $y_{n+1} = (1 - \lambda_n)\widetilde{y}_{n+1} + \lambda_n\widehat{y}_{n+1}$, it seems also desirable that $g(\widetilde{y}_{n+1})$ and $g(\widehat{y}_{n+1})$ have opposite signs in order to get a real value λ_n at each step satisfying $g(y_{n+1}) = 0$. To estimate $g(\widehat{y}_{n+1})$, if $\widehat{y}_{n+1}$ has order 1, we can write

$$g(\widehat{y}_{n+1}) = g(\widetilde{y}_{n+1}) + \nabla g(\widetilde{y}_{n+1})(\widehat{y}_{n+1} - \widetilde{y}_{n+1}) + O(h^4),$$

where the two first terms in the right hand side are, respectively, $O(h^4)$ and $O(h^2)$. So, we can approximate

$$g(\widehat{y}_{n+1}) \approx g(\widetilde{y}_{n+1}) + \nabla g(\widetilde{y}_{n+1})(\widehat{y}_{n+1} - \widetilde{y}_{n+1}).$$

Therefore, if possible, we will look for values $\widehat{y}_{n+1}$ satisfying

$$g(\widetilde{y}_{n+1})\Big[g(\widetilde{y}_{n+1}) + h\sum_{i=1}^{3}(\widehat{b}_i - \widetilde{b}_i)k_i\Big] < 0, \tag{2.4}$$

where $k_i = \nabla g(\widetilde{y}_{n+1})g_i$, and g_i are the stages of the RK pair (2.2):

$$g_i = f(y_n + h\sum_{j=1}^{i-1} a_{ij}g_j), \quad i = 1, 2, 3.$$

In relation with our first criterion (2.3), we have:

Theorem 1. *Let us consider the projection method* (1.2) *where* $\widetilde{y}_{n+1}$ *and* $\widehat{y}_{n+1}$ *represent the RK methods of the embedded pair given in* (2.2)*. Then,*

i) $\widehat{y}_{n+1}$ *satisfies* (2.3) *with dispersion order 2* $\Leftrightarrow \widehat{b}_2 < -\frac{1}{3} + 3\widehat{b}_1$.

ii) $\widehat{y}_{n+1}$ *satisfies* (2.3) *with dispersion order 4* $\Leftrightarrow \widehat{b}_2 = -\frac{1}{3} + 3\widehat{b}_1,\ \widehat{b}_1 > \frac{13}{45}$.

iii) $\widehat{y}_{n+1}$ *satisfies* (2.3) *with dispersion order 6* $\Leftrightarrow \widehat{b}_1 = \frac{13}{45},\ \widehat{b}_2 = \frac{8}{15}$.

Proof. The dispersion errors for $\widetilde{y}_{n+1}$ and $\widehat{y}_{n+1}$ turn out to be, respectively,

$$\widetilde{\phi}(\nu) = -\frac{1}{30}\nu^5 + O(\nu^7), \quad \widehat{\phi}(\nu) = \frac{1}{24}(-1 + 9\widehat{b}_1 - 3\widehat{b}_2)\nu^3 + O(\nu^5),$$

from which item *i*) is deduced. Clearly, $\widehat{b}_2 = -\frac{1}{3} + 3\widehat{b}_1$ makes zero the coefficient of ν^3 in $\widehat{\phi}(\nu)$, and then we obtain

$$\widehat{\phi}(\nu) = \frac{1}{90}(-13 + 45\widehat{b}_1)\nu^5 + O(\nu^7),$$

which gives rise to item *ii*). Finally, cancelling the principal error term of $\widehat{\phi}(\nu)$, we obtain

$$\widehat{\phi}(\nu) = \frac{1}{1575}\nu^7 + O(\nu^9),$$

and item *iii*) is proved. □

According this result, we still have some degrees of freedom after choosing the coefficients $\widehat{b}_1, \widehat{b}_2$ satisfying (2.3) in the cases *i*) and *ii*). In both situations, we wonder if the free parameters can be chosen so that the condition (2.4) is also satisfied. Thus, the following result studies when the combination of Theorem 1, item *ii*), with the criterion (2.4) is possible.

Theorem 2. *The approximation* $\widehat{y}_{n+1}$ *satisfies* (2.3) *with dispersion order 4 and also* (2.4) *if and only if*

$$\begin{cases} \textit{sign}\ g(\widetilde{y}_{n+1}) = -\textit{sign}\ (k_1 + 3k_2 - 4k_3), \\ \widehat{b}_1 > \max\{\frac{13}{45}, \gamma\}, \quad \widehat{b}_2 = -\frac{1}{3} + 3\widehat{b}_1. \end{cases}$$

where

$$\gamma = \frac{2}{9} - \frac{g(\widetilde{y}_{n+1})}{h(k_1 + 3k_2 - 4k_3)}, \quad (k_1 + 3k_2 - 4k_3 \neq 0).$$

Proof. After substituting $\widehat{b}_2 = -\frac{1}{3} + 3\widehat{b}_1$ and the coefficients $\widetilde{b}_i$, $i = 1, 2, 3$, into the approximation to $g(\widehat{y}_{n+1})$, it results

$$g(\widehat{y}_{n+1}) \approx g(\widetilde{y}_{n+1}) + h(k_1 + 3k_2 - 4k_3)\left(\widehat{b}_1 - \frac{2}{9}\right).$$

If $k_1 + 3k_2 - 4k_3 = 0$, clearly (2.4) can not be satisfied, and the same happens if sign $g(\widetilde{y}_{n+1}) =$ sign $(k_1 + 3k_2 - 4k_3)$ since $\widehat{b}_1 > \frac{13}{45} > \frac{2}{9}$. If sign $g(\widetilde{y}_{n+1}) = -$sign $(k_1 + 3k_2 - 4k_3)$, the condition (2.4) leads to $\widehat{b}_1 > \gamma$, and the proof is completed. □

The combination of (2.3) in the form *i*) established in Theorem 1 together with (2.4) has been studied in [6], recently submitted to publication by the authors. Here is the result obtained:

Theorem 3. *The RK method* $\widehat{y}_{n+1}$ *with coefficients* $(A, \widehat{b}^T)$ *given in* (2.2) *satisfies* (2.3) *with dispersion order 2 and also* (2.4) *if and only if one of these four situations happens:*

a)

$$\begin{cases} \textit{sign } g(\widetilde{y}_{n+1}) = \textit{sign } (k_2 - k_3), \\ \widehat{b}_1 \textit{ arbitrary}, \ \widehat{b}_2 < \min\left\{ -\frac{1}{3} + 3\widehat{b}_1, \alpha(\widehat{b}_1) \right\}. \end{cases}$$

b)

$$\begin{cases} \textit{sign } g(\widetilde{y}_{n+1}) = -\textit{sign } (k_2 - k_3), \\ \widehat{b}_1 \begin{cases} < \gamma, & \textit{if sign } (k_2 - k_3) = -\textit{sign } (k_1 + 3k_2 - 4k_3), \\ > \gamma, & \textit{if sign } (k_2 - k_3) = \textit{sign } (k_1 + 3k_2 - 4k_3), \end{cases} \\ \widehat{b}_2 \ni \alpha(\widehat{b}_1) < \widehat{b}_2 < -\frac{1}{3} + 3\widehat{b}_1. \end{cases}$$

c)

$$\begin{cases} k_2 = k_3, \quad \textit{sign } g(\widetilde{y}_{n+1}) = \textit{sign } (k_1 - k_3), \\ \widehat{b}_2 < -\frac{1}{3} + 3\beta, \ \widehat{b}_1 \ni \frac{1}{9} + \frac{\widehat{b}_2}{3} < \widehat{b}_1 < \beta. \end{cases}$$

d)

$$\begin{cases} k_2 = k_3, \ \textit{sign } g(\widetilde{y}_{n+1}) = -\textit{sign } (k_1 - k_3), \\ \widehat{b}_2 \textit{ arbitrary}, \ \widehat{b}_1 > \max\{\beta, \frac{1}{9} + \frac{\widehat{b}_2}{3}\}. \end{cases}$$

where

$$\alpha(\widehat{b}_1) = \frac{k_3 - k_1}{k_2 - k_3}\widehat{b}_1 + \frac{2k_1 + 3k_2 - 5k_3}{9(k_2 - k_3)} - \frac{g(\widetilde{y}_{n+1})}{h(k_2 - k_3)}, \quad (k_2 \neq k_3),$$

$$\beta = \frac{2}{9} - \frac{g(\widetilde{y}_{n+1})}{h(k_1 - k_3)}, \quad (k_1 \neq k_3).$$

□

Let us notice that this theorem assures the existence of coefficients $\widehat{b_i}$, $i = 1, 2$, satisfying both criteria, (2.3) and (2.4), almost in any situation.

According to the previous results, we have designed an algorithm to compute, at each step, the coefficients $\widehat{b_1}$ and $\widehat{b_2}$. When it is possible, it chooses them so that both criteria (2.3) and (2.4) are satisfied with the highest dispersion order for $\widehat{y}_{n+1}$. In other case, it takes the coefficients satisfying (2.3) according to Theorem 1, item *iii*). More specifically, it proceeds as follows:

1. Calculate $\widehat{y}_{n+1}$ from $\widehat{b_1} = \frac{13}{45}$, $\widehat{b_2} = \frac{8}{15}$. If $g(\widehat{y}_{n+1})g(\widetilde{y}_{n+1}) < 0$, then we take those values for the coefficients.
2. If sign $g(\widetilde{y}_{n+1}) = -\text{sign}\,(k_1 + 3k_2 - 4k_3)$, then we take:
$$\widehat{b_1} = \max\{\frac{13}{45}, \gamma\} + 0.1, \quad \widehat{b_2} = -\frac{1}{3} + 3\widehat{b_1}.$$
3. If sign $g(\widetilde{y}_{n+1}) = \text{sign}\,(k_2 - k_3)$, then we take:
$$\widehat{b_1} = 0, \quad \widehat{b_2} = \min\Big\{-\frac{1}{3} + 3\widehat{b_1}, \alpha(\widehat{b_1})\Big\} - 0.1.$$
4. If sign $g(\widetilde{y}_{n+1}) = -\text{sign}\,(k_2 - k_3) = \text{sign}\,(k_1 + 3k_2 - 4k_3)$, then we take:
$$\widehat{b_1} = \gamma - 0.1, \quad \widehat{b_2} = -\frac{1}{6} + \frac{3}{2}\widehat{b_1} + \frac{\alpha(\widehat{b_1})}{2}.$$
5. If $k_2 = k_3$ and sign $g(\widetilde{y}_{n+1}) = \text{sign}\,(k_1 - k_3)$, then we take:
$$\widehat{b_2} = -\frac{1}{3} + 3\beta - 0.1, \quad \widehat{b_1} = \beta - \frac{0.1}{6}.$$
6. In any other case we take:
$$\widehat{b_1} = \frac{13}{45}, \quad \widehat{b_2} = \frac{8}{15}.$$

§3. Numerical experiments

We are going to check the projection technique developed in the previous section by applying the algorithm shown there to obtain appropriate coefficients $\widehat{b_1}$ and $\widehat{b_2}$ ($\widehat{b_3} = 1 - \widehat{b_1} - \widehat{b_2}$). From those values we obtain the RK approximation $\widehat{y}_{n+1}$, and then, we compute the direction of the projection w_n according to (2.1). Newton iteration has been carried out to obtain in each step the parameter λ_n. Finally, we obtain y_{n+1}, projection of the Bogacki–Shampine method $\widetilde{y}_{n+1}$ according to (1.2). This projected method y_{n+1} will be denoted by *pBS*3. We will compare it with the own Bogacki–Shampine method, which will be denoted by *BS*3, and also with the standard projection of *BS*3, denoted by *pstBS*3, which takes $w_n = \nabla g(\widetilde{y}_{n+1})$ [7, pp. 106]. All the integrations have been carried out with fixed step size.

Our first test problem is the known as Euler problem [7, pp. 95]:

$$\begin{aligned} y_1'(t) &= (c_3 - c_2)y_2y_3, \\ y_2'(t) &= (c_1 - c_3)y_1y_3, \\ y_3'(t) &= (c_2 - c_1)y_1y_2. \end{aligned}$$

We have taken $c_1 = 1/0.345$, $c_2 = 1/0.653$ and $c_3 = 1$, and the initial conditions $y_1(0) = 0.5$, $y_2(0) = 0.2$, $y_3(0) = \sqrt{1 - 0.5^2 - 0.2^2}$. Its solution is periodic and its period T depends on the two first integrals:

$$E(y_1, y_2, y_3) = (c_1y_1^2 + c_2y_2^2 + c_3y_3^2)/2, \quad L^2(y_1, y_2, y_3) = y_1^2 + y_2^2 + y_3^2.$$

For this problem the projections have been done so that the numerical solution preserves just the period T.

In Figure 1 we have represented the euclidean norm of the global error at the end of each integration interval against the number of periods in a log-log scale. As it can be seen, the best results correspond to *pBS*3, the method projected according the technique studied in this paper. It behaves even better than the projection method *pstBS*3, which uses the standard projection technique. The two dash-dotted straight lines with slopes 1 and 2 indicate that the global error grows linearly with the number of periods for the two projection methods, whereas the growth is quadratic for the basic formula *BS*3.

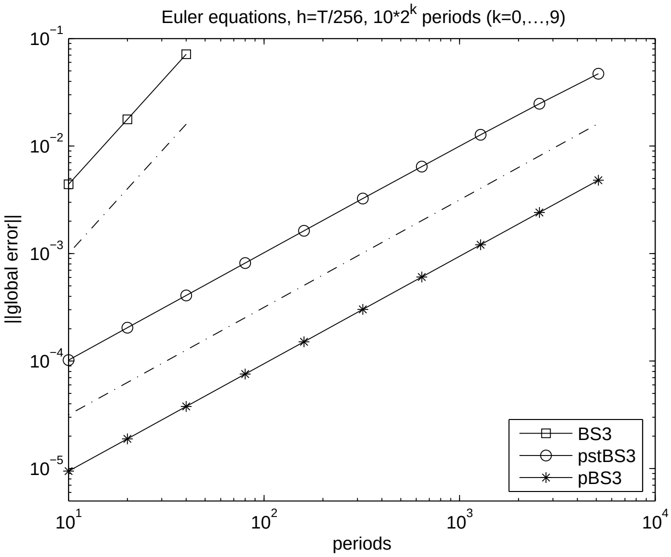

Figure 1: Euler equations, global error against periods, log-log scale.

We also present here analogous numerical experiments for the Lotka–Volterra problem, given by [7, pp. 229]:

$$u'(t) = u(v - 2), \quad v'(t) = v(1 - u),$$

with initial conditions $u(0) = v(0) = 1$. Now, the projections preserve the function

$$H(u, v) = u - \ln u + v - 2 \ln v,$$

which is a first integral of that differential system. Figure 2 shows that the new projection method gives rise to lower global errors than the other two methods, and similar comments to those of Figure 1 can be done.

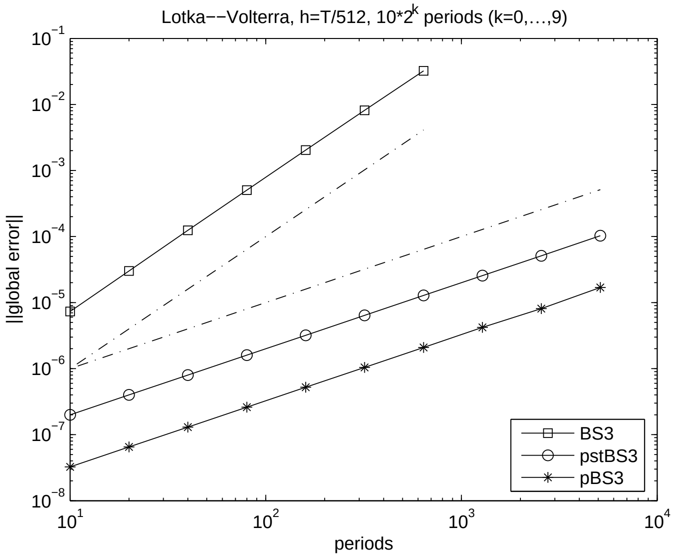

Figure 2: Lotka–Volterra, global error against periods, log-log scale.

Acknowledgements

This work was supported by D.G.I. project MTM2010-21630-C02.

References

[1] Bogacki, P., and Shampine, L. F. A 3(2) pair of Runge-Kutta formulas. *Applied Mathematics Letters 2*, 4 (1989), 321–325.

[2] Calvo, M., Franco, J. M., Montijano, J. I., and Rández, L. Explicit Runge-Kutta methods for initial value problems with oscillating solutions. *J. Comp. Appl. Math. 76* (1996), 195–212.

[3] Calvo, M., Hernández-Abreu, D., Montijano, J. I., and Rández, L. On the preservation of invariants by explicit Runge-Kutta methods. *SIAM J. Sci. Comput. 28*, 3 (2006), 868–885.

[4] Calvo, M., Laburta, M. P., Montijano, J. I., and Rández, L. Approximate preservation of quadratic first integrals by explicit Runge-Kutta methods. *Adv. Comput. Math. 32*, 3 (2010), 255–274.

[5] Calvo, M., Laburta, M. P., Montijano, J. I., and Rández, L. Projection methods preserving Lyapunov functions. *BIT Numer. Math. 50*, 2 (2010), 223–241.

[6] Calvo, M., Laburta, M. P., Montijano, J. I., and Rández, L. Runge-Kutta projection methods with low dispersion and dissipation errors. *submitted to publication* (2013).

[7] Hairer, E., Lubich, C., and Wanner, G. *Geometric Numerical Integration*. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 2002. Structure-preserving algorithms for ordinary differential equations.

[8] van der Houwen, P. J., and Sommeijer, B. P. Explicit Runge-Kutta (-Nyström) methods with reduced phase errors for computing oscillating solutions. *SIAM J. Numer. Anal. 24*, 3 (1987), 595–617.

María Pilar Laburta
IUMA, Departamento de Matemática Aplicada
Escuela de Ingeniería y Arquitectura. Universidad de Zaragoza
50018-Zaragoza, Spain
laburta@unizar.es

Juan Ignacio Montijano
IUMA, Departamento de Matemática Aplicada
Facultad de Ciencias. Universidad de Zaragoza
50009-Zaragoza, Spain
monti@unizar.es