SIAM J. SCI. COMPUT.
Vol. 28, No. 3, pp. 868–885

ON THE PRESERVATION OF INVARIANTS BY EXPLICIT RUNGE–KUTTA METHODS*

M. CALVO†, D. HERNÁNDEZ-ABREU‡, J. I. MONTIJANO†, AND L. RÁNDEZ†

Abstract. A new strategy to preserve invariants in the numerical integration of initial value problems with explicit Runge–Kutta methods is presented. It is proved that this technique retains the order of the original method, has an easy and cheap implementation, and can be used in adaptive Runge–Kutta codes. Some numerical experiments with the classical code of Dormand and Prince, DoPri5(4), based on a pair of embedded methods with orders 5 and 4, are presented to show the behavior of the new method for several problems which possess invariants.

Key words. initial value problems, explicit Runge–Kutta methods, numerical geometric integration, preservation of invariants, variable step-size codes

AMS subject classifications. 65L05, 65L06

DOI. 10.1137/04061979X

1. Introduction and basic definitions.

First integrals and invariants of (autonomous) differential systems

$$y' = f(y), \tag{1}$$

where $f : \mathcal{D} \subset \mathbb{R}^m \to \mathbb{R}^m$ is a sufficiently smooth function, play an important role both in qualitative and quantitative studies of the flow of (1). These invariants allow us to describe the geometry of the orbits and also to check the accuracy of numerical integrators of the corresponding differential equation.

A function $G = G(y) : \widehat{\mathcal{D}} \subset \mathbb{R}^m \to \mathbb{R}^l$ of class $\mathcal{C}^1(\widehat{\mathcal{D}})$, $\widehat{\mathcal{D}} \subset \mathcal{D}$, is called an l-invariant system of (1) in $\widehat{\mathcal{D}}$ (see [15, p. 61]) if

(i) there exists some $y^0 \in \widehat{\mathcal{D}}$ such that $G(y^0) = 0$,
(ii) for all solutions $y = y(t)$ of (1), $G(y(t)) = 0$ holds either for every t or for no t in the interval of definition of $y(t)$.

In such a case we may consider the flow defined by (1) restricted to the $(m-l)$-dimensional manifold $\mathcal{M} = \{y \in \widehat{\mathcal{D}}; G(y) = 0\}$ as a differential equation on $\mathcal{M}$. In the case $l = 1$, the invariant system is called an invariant relation. In connection with the above concept, it is worth remarking that for $l > 1$ the l scalar relations which constitute the invariant system need not be invariant relations. An l-invariant system is also called a weak invariant (see [5]), and it is not difficult to see that $G(y)$ is a weak invariant of (1) if and only if there exists some $y^0 \in \widehat{\mathcal{D}}$ such that $G(y^0) = 0$, and $\nabla G(y)^T f(y) = 0$ for all $y \in \mathcal{M}$.

On the other hand, a scalar function $F : \widehat{\mathcal{D}} \subset \mathbb{R}^m \to \mathbb{R}$ of class $C^1(\widehat{\mathcal{D}}), \widehat{\mathcal{D}} \subset \mathcal{D}$, is a first integral (or a conserved quantity) of (1) if $F(y(t))$ is a constant along any

*Received by the editors November 29, 2004; accepted for publication (in revised form) December 27, 2005; published electronically June 9, 2006.
http://www.siam.org/journals/sisc/28-3/61979.html

†Departamento Matemática Aplicada, Universidad de Zaragoza, 50009-Zaragoza, Spain (calvo@unizar.es, monti@unizar.es, randez@unizar.es). The work of these authors was supported by project MTM2004-06466-C02-01.

‡Departamento de Análisis Matemático, Universidad de La Laguna, 38271, La Laguna, Spain (dhabreu@ull.es). The work of this author was supported by project MTM2004-06466-C02-02 and grant AP2002-2761.

solution path $y = y(t)$ of (1). In such a case, since

$$0 = \frac{dF(y(t))}{dt} = \nabla F(y(t))^T \cdot f(y(t)) \qquad \text{for all } t,$$

it follows that $\nabla F(y)^T \cdot f(y) = 0$ for all $y \in \widehat{\mathcal{D}}$. Now, for all $y^0 \in \widehat{\mathcal{D}}$ the solution $y = y(t)$ of (1) such that $y(0) = y^0$ is contained in the hypersurface $\mathcal{M}_{y^0} = \{y \in \mathbb{R}^m \,;\, F(y) = F(y^0)\}$. First integrals are also called strong invariants (see [5, 7]) and quadratic integrals $G(y) = y^T S y$, where $S \in \mathbb{R}^{m\times m}$ is a constant symmetric matrix, are strong invariants of (1) if and only if $y^T\, S\, f(y) = 0$ for all $y \in \mathcal{D}$.

In order to simplify the presentation, we will use in the following the term invariant of the differential system (1) for either a first integral or an invariant system (or relation) of (1).

Since many general numerical integrators do not preserve first integrals or invariants of their corresponding differential systems, these functions have been widely used as a tool to measure the accuracy and the long-term behavior of these integrators. As a matter of fact, in the last two decades there has been an increasing interest in numerical integrators that preserve as far as possible the qualitative properties of the underlying differential systems. In particular, a natural requirement is to consider those numerical Runge–Kutta (RK) methods that preserve first integrals or invariants. It was proved by Cooper [3] that all RK methods conserve linear invariants. However, an irreducible RK method preserves all quadratic invariants $G(y) = y^T S y + \mu^T y + \nu$ with constants $S \in \mathbb{R}^{m\times m}, \mu \in \mathbb{R}^m, \nu \in \mathbb{R}$, if and only if their coefficients $(A \in \mathbb{R}^{s\times s}, b \in \mathbb{R}^s)$ satisfy $m_{ij} \equiv b_i a_{ij} + b_j a_{ji} - b_i b_j = 0$, $1 \le i \le j \le s$. These conditions, which also imply that (A, b) is a symplectic method, pose strong requirements on the coefficients of the method which are satisfied only by some implicit RK methods and have been extensively studied in connection with symplectic methods. Moreover, it can be seen (see, e.g., [5, Thm. 3.3, p. 102]) that for $n \ge 3$ no RK method can conserve all polynomial invariants of degree n. Further studies on the preservation of invariants by numerical methods have been carried out in [7, 9, 11].

An alternative attempt to preserve invariants in numerical integrations is proposed by Schropp [10]. This author modifies the original flow of (1) by stabilizing the invariant set $\mathcal{M}_{y^0} = \{y \in \mathbb{R}^m \,;\, G(y) = G(y^0)\}$ and considers the modified problem

$$y' = f(y) - \nabla G(y)\; a(y)(G(y) - G(y^0)),$$

where $a(y)$ is a suitably chosen smooth function so that $\mathcal{M}_{y^0}$ is an exponentially attracting set for the modified problem. In view of this, the numerical solutions generated by the method are expected to approach this attracting set; however, the modified problem may become very stiff depending on the choice of $a(y)$.

In view of the above limitations several techniques have been proposed to preserve invariants in practical numerical integrators. Thus, del Buono and Mastroserio in [1], starting with an explicit s-stage RK method $(A = (a_{ij}), b = (b_i))$ given by the equations

$$\varphi_h(y_n) = y_n + h \sum_{j=1}^{s} b_j f(Y_{n,j}), \tag{2}$$

with

$$Y_{n,1} = y_n, \quad Y_{n,i} = y_n + h\sum_{j=1}^{i-1} a_{ij} f(Y_{n,j}) \quad (2 \le i \le s), \tag{3}$$

modify the advancing solution (2) by introducing at each step a scalar $\gamma_n = \gamma_n(y_n, h)$, so that the new method $\widehat{\varphi}_h$ defined by (3) and

$$\widehat{\varphi}_h(y_n) = y_n + h\gamma_n \sum_{j=1}^{s} b_j f(Y_{n,j}) \tag{4}$$

preserves a given invariant of (1). Clearly, this requirement implies that γ_n will be a problem-dependent scalar such that $\gamma_n \to 1$ when $h \to 0$ and may vary from step to step. These authors [1] have proved that for all differential systems (1) possessing a quadratic invariant $G(y) = y^T S y$ with a symmetric S, there exist fourth-order explicit RK methods φ_h with four stages such that the modified methods $\widehat{\varphi}_h$ given by (3) and (4) preserve the quadratic invariant and have order three for a suitable value γ_n. Further, $\widehat{\varphi}_h$ attains order four at $t_n + h\gamma_n$, and therefore $\widehat{\varphi}_h$ would get (global) order four at the nonuniform grid $\{\widehat{t}_n\}_{n=0}^N$ with $\widehat{t}_0 = t_0$ and $\widehat{t}_{n+1} = \widehat{t}_n + h\gamma_n$. To compute γ_n at each step, the authors [1] use the formula $\gamma_n = 1 - (\delta_n/\eta_n)$ with

$$\eta_n = \sum_{i,j=1}^{4} b_i b_j \langle f(Y_{n,i}), f(Y_{n,j})\rangle, \quad \delta_n = \eta_n - 2\sum_{i=2}^{4}\sum_{j=1}^{i-1} b_i a_{ij} \langle f(Y_{n,i}), f(Y_{n,j})\rangle,$$

where $\langle u, v\rangle \equiv u^T S v$ for $u, v \in \mathbb{R}^m$ is the bilinear form associated to the invariant $G(y) = y^T S y$. This means that the computational cost of $\widehat{\varphi}_h$ at each step includes $s(s-1)/2 = 6$ bilinear forms $\langle f(Y_{n,i}), f(Y_{n,j})\rangle$, apart from the function evaluations of (3). Note that $\widehat{\varphi}_h(y_n) = \varphi_h(y_n) + (\gamma_n - 1)(\varphi_h(y_n) - y_n)$ can be considered as the projection of $\varphi_h(y_n)$ onto the invariant manifold $\mathcal{M}_{y_n} = \{y \ ; \ G(y) = G(y_n)\}$ along the direction $\varphi_h(y_n) - y_n$. In the remainder of this paper this technique will be called the incremental direction technique.

An alternative technique (see [5, p. 106]), known as the standard projection method, consists of combining a standard RK method φ_h (or any one-step method) together with an orthogonal projection P onto the invariant manifold $\mathcal{M}_{y_n}$ at each step $(t_n, y_n) \to (t_{n+1} = t_n + h, y_{n+1})$. For the Euclidean norm, it would be given by $y_{n+1} = (P \cdot \varphi_h)(y_n)$, with y_{n+1} defined by $y_{n+1} = \varphi_h(y_n) + \nabla G(y_{n+1})\ \lambda_n$, with $G(y_{n+1}) = G(y_n)$. Due to the implicitness, this system is replaced in [5] by

$$y_{n+1} = \varphi_h(y_n) + \nabla G(\varphi_h(y_n))\ \lambda_n \quad \text{with} \quad G(y_{n+1}) = G(y_n), \tag{5}$$

and the calculation of λ_n at each step is carried out by applying a simplified Newton iteration to the corresponding nonlinear equation. As remarked in [5], since λ_n is small, taking the starting value $\lambda_{n,0} = 0$, one simplified Newton iteration is usually enough for practical purposes. It can be proved that (5) preserves the order of the original method $\varphi_h(y_n)$ but requires us to provide an analytical expression of the gradient $\nabla G(y)$ that will be evaluated at each step and also will be used to obtain λ_n by a simplified Newton iterative scheme in the solution of the implicit equations $G(y_{n+1}) = G(y_n)$. On the other hand, although RK methods preserve linear invariants and are affine invariant, the standard projection method (5) in general does

not preserve these properties, and this fact may be inconvenient for the qualitative behavior of the integrator.

As follows from (4) and (5), the above projection techniques can be included in the general form

$$\widehat{\varphi}_h(y_n) = \varphi_h(y_n) - \lambda_n \Phi_n, \tag{6}$$

where the vector $\Phi_n \in \mathbb{R}^m$ defines the direction of projection and λ_n is a scalar chosen so that $\widehat{\varphi}_h(y_n)$ belongs to the invariant manifold. In the case (4) the direction $\Phi_n = \varphi_h(y_n) - y_n$ is defined by the original method and a zero-order method, whereas in the case (5) of the simplified projection, Φ_n is orthogonal to the invariant at $\varphi_h(y_n)$. Clearly, other possible directions Φ_n may be considered. In particular, taking into account that for many explicit RK methods several low order embedded approximations of type

$$\widetilde{\varphi}_h(y_n) = y_n + h \sum_{j=1}^{s} \widetilde{b}_j f(Y_{n,j}) \tag{7}$$

are usually available without additional cost, one is tempted to take $\Phi_n = \varphi_h(y_n) - \widetilde{\varphi}_h(y_n)$, and then the resulting projected method (6)

$$\widehat{\varphi}_h(y_n) = y_n + h \sum_{j=1}^{s} ((1-\lambda_n) b_j + \lambda_n \widetilde{b}_j) f(Y_{n,j}) \tag{8}$$

has the form of an RK method with (variable) weights $\widehat{b}_j = (1-\lambda_n) b_j + \lambda_n \widetilde{b}_j$ and retains important properties of these methods, such as the preservation of linear invariants, the affine invariance, and a low computational cost.

The aim of this paper is to study projected explicit RK methods of type (8) that preserve general invariants with the purpose of being used in numerical integrators based on adaptive RK methods. In section 2, a revision of the case $\widetilde{\varphi}_h(y_n) = y_n$, in which the companion method has order zero, is presented with the purpose of generalizing the results given in [1] and also of showing a more efficient way of computing λ_n. In section 3, a general study for quadratic invariants and embedded methods $\widetilde{\varphi}_h$ with order q $(1 \le q < p)$ is carried out. In section 4 this study is generalized to nonquadratic scalar invariants and to the case of several scalar invariants. Further, some remarks on how to include the proposed schemes in adaptive RK codes are also given. Finally, in section 5, some numerical experiments with the classical embedded pair of explicit RK methods of orders five and four, DoPri5(4) [4], are shown in order to make clear how our technique performs on several test problems.

2. The incremental direction technique revisited. In [1] del Buono and Mastroserio propose to advance the integration $t_n \to t_{n+1} = t_n + h$ of (1) with the RK method $\widehat{\varphi}_h$ given by (4), where γ_n is properly chosen so that the method preserves a fixed quadratic invariant $G(y) = y^T\, S\, y = \langle y, y \rangle$. Here we will write $\widehat{\varphi}_h$ in the equivalent form

$$\widehat{\varphi}_h(y_n) = \varphi_h(y_n) - \lambda_n(\varphi_h(y_n) - y_n) = (1-\lambda_n)\varphi_h(y_n) + \lambda_n y_n, \tag{9}$$

so that the scalar $\lambda_n = 1 - \gamma_n$ will be a small quantity at each step.

Next it will be seen that for all RK methods φ_h with order $p \ge 2$ the projected method $\widehat{\varphi}_h$ has order $p - 1$. This result generalizes Proposition 4 of [1] in two

directions: First, it extends the statement on fourth-order RK methods to any order $p \geq 2$; second, there are no restrictions on the coefficients of the method φ_h in contrast with the situation in [1]. Finally, we will show that for any order p, λ_n (or γ_n) can be computed with two bilinear form evaluations $\langle \cdot, \cdot \rangle$ per step.

With the notation (9) and putting $\varphi_h(y_n) - y_n = h\Delta_n$, the preservation of the quadratic invariant $G(y) = y^T S y = \langle y, y \rangle$ holds if and only if λ_n satisfies

$$2\langle y_n, h\Delta_n \rangle + (1 - \lambda_n)\langle h\Delta_n, h\Delta_n \rangle = 0. \tag{10}$$

If $\langle y_n, h\Delta_n \rangle = \langle h\Delta_n, h\Delta_n \rangle = 0$, then $\langle \widehat{\varphi}_h(y_n), \widehat{\varphi}_h(y_n) \rangle = \langle y_n, y_n \rangle$ for all λ_n, and therefore it would be enough to take $\lambda_n = 0$ (i.e., the original method φ_h) to preserve the invariant. If either $\langle y_n, h\Delta_n \rangle = 0$ or else $\langle h\Delta_n, h\Delta_n \rangle = 0$, equation (10) cannot be satisfied by $\lambda_n \neq 1$ (note that $\lambda_n = 1$ leads to the method $\widehat{\varphi}_h(y_n) = y_n$ producing a constant output) and therefore it is not possible to preserve the invariant. Hence the condition

$$\langle y_n, \Delta_n \rangle \, \langle \Delta_n, \Delta_n \rangle \neq 0 \tag{11}$$

must be satisfied in order to make (10) solvable. Observe that if φ_h is Euler's method, since $\Delta_n = f(y_n)$, condition (11) does not hold. Thus, only methods with order $p \geq 2$ will be considered.

THEOREM 2.1. *Let φ_h be a method with order $p \geq 2$ and suppose that* (11) *holds.*

(i) *The projected method $\widehat{\varphi}_h$ given by* (9) *preserves the quadratic invariant $G(y) = \langle y, y \rangle$ provided that λ_n is chosen as*

$$\lambda_n = \frac{\langle 2y_n + h\Delta_n, h\Delta_n \rangle}{\langle h\Delta_n, h\Delta_n \rangle} = \frac{\langle \varphi_h(y_n), \varphi_h(y_n) \rangle - \langle y_n, y_n \rangle}{\langle \varphi_h(y_n) - y_n, \varphi_h(y_n) - y_n \rangle}. \tag{12}$$

(ii) *If $\langle f(y_n), f(y_n) \rangle \neq 0$, then the projected method $\widehat{\varphi}_h$ has order $p - 1$.*

Proof. Under the assumption (11), equation (10) has the solution λ_n given by (12), and therefore the method $\widehat{\varphi}_h$ preserves the quadratic invariant. On the other hand, since φ_h has order p, we have

$$\varphi_h(y_n) = y_n + h\Delta_n = y(t_n + h) + \epsilon_{n+1},$$

where $\epsilon_{n+1} = C_{p+1}h^{p+1} + \mathcal{O}(h^{p+2})$ is the local error of φ_h at y_n with step size h and $y(t_n + t)$ is the local solution of (1) at (t_n, y_n); i.e., it satisfies $y(t_n) = y_n$ and

$$\langle y(t_n + h), y(t_n + h) \rangle = \langle y(t_n), y(t_n) \rangle = \langle y_n, y_n \rangle$$

for all h. Then, for the numerator and denominator of λ_n in (12) we have

$$\begin{aligned} \langle \varphi_h(y_n), \varphi_h(y_n) \rangle - \langle y_n, y_n \rangle &= 2\langle y_n, C_{p+1} \rangle h^{p+1} + \mathcal{O}(h^{p+2}), \\ \langle h\Delta_n, h\Delta_n \rangle &= h^2 \langle y'(t_n), y'(t_n) \rangle + \mathcal{O}(h^3), \end{aligned}$$

and λ_n behaves like $O(h^{p-1})$ providing that $\langle y'(t_n), y'(t_n) \rangle = \langle f(y_n), f(y_n) \rangle \neq 0$, and, according to (9), this implies that $\widehat{\varphi}_h$ is a method of order $p - 1$. ∎

To end this section, observe that it follows from (12) that the computation of λ_n requires only two bilinear forms per step in contrast with [1] which needs six bilinear forms for $s = 4$. Moreover, since the two required bilinear forms are $\langle \varphi_h(y_n), \varphi_h(y_n) \rangle, \langle \varphi_h(y_n), y_n \rangle$, only one matrix-vector product is necessary.

3. A general projection technique for quadratic invariants. Let φ_h be an explicit s-stage RK method (A, b) with order p defined by equations (2)–(3), and let $\widetilde{\varphi}_h$ be an embedded method (7) with order q $(1 \le q < p)$ defined by the weights $\widetilde{b}_i,\ i = 1, \dots, s$. As remarked above, practical RK methods possess embedded methods with several orders $(1 \le q \le p-1)$, so that we have wide freedom in the choice of $\widetilde{\varphi}_h$ and the computation of the embedded method does not require additional function evaluations. Suppose that (1) possesses a quadratic invariant $G(y) = y^T S y = \langle y, y\rangle$. We propose to complete the step $t_n \to t_{n+1}$, projecting $\varphi_h(y_n)$ onto the manifold defined by the invariant $\mathcal{M}_{y_n} = \{y \in \mathbb{R}^m; \langle y, y\rangle = \langle y_n, y_n\rangle\}$ along the direction given by the unit vector

$$w_n = w(y_n, h) = \frac{\varphi_h(y_n) - \widetilde{\varphi}_h(y_n)}{\|\varphi_h(y_n) - \widetilde{\varphi}_h(y_n)\|_2}, \tag{13}$$

where $\|\cdot\|_2$ is the Euclidean norm.

Then the projected method $\widehat{\varphi}_h$ is given by

$$\widehat{\varphi}_h(y_n) = \varphi_h(y_n) - \lambda_n w_n, \tag{14}$$

with the scalar $\lambda_n = \lambda(y_n, h)$ such that

$$G(\widehat{\varphi}_h(y_n)) - G(y_n) \equiv \langle \widehat{\varphi}_h(y_n), \widehat{\varphi}_h(y_n)\rangle - \langle y_n, y_n\rangle = 0. \tag{15}$$

Substituting (14) into (15), λ_n must satisfy

$$\alpha_n \lambda_n^2 - 2\beta_n \lambda_n + \delta_n = 0, \tag{16}$$

with

$$\begin{aligned} \alpha_n &= \alpha(y_n, h) = \langle w_n, w_n\rangle, \\ \beta_n &= \beta(y_n, h) = \langle \varphi_h(y_n), w_n\rangle, \\ \delta_n &= \delta(y_n, h) = \langle \varphi_h(y_n), \varphi_h(y_n)\rangle - \langle y_n, y_n\rangle. \end{aligned} \tag{17}$$

Given $h > 0$ and y_n, if $\delta(y_n, h) = 0$, the original method $\varphi_h(y_n)$ preserves the quadratic invariant in this step and in such a case we would take $\lambda_n = 0$.

If $\delta_n \neq 0$, let us suppose $\beta_n^2 - \alpha_n \delta_n > 0$; then we take for λ_n the real root of (16) closest to $\lambda = 0$ given by

$$\lambda_n = \begin{cases} \dfrac{\delta_n}{\beta_n + \operatorname{sign}\beta_n \sqrt{\beta_n^2 - \alpha_n \delta_n}} & \text{if} \quad \beta_n \neq 0, \\ (-\delta_n/\alpha_n)^{1/2} & \text{if} \quad \beta_n = 0. \end{cases} \tag{18}$$

To study the order of the projected method (14), (15), observe that denoting by $C_{q+1}(y_n)h^{q+1}$ the leading term of the local error of the embedded method $\widetilde{\varphi}_h(y_n)$, $w_n = C_{q+1}/\|C_{q+1}\|_2 + \mathcal{O}(h)$ and then

$$\beta(y_n, 0) = \langle \varphi_h(y_n), w(y_n, h)\rangle|_{h=0} = \left\langle y_n, \frac{C_{q+1}}{\|C_{q+1}\|_2} \right\rangle. \tag{19}$$

Now we have the following theorem.

THEOREM 3.1. *Let φ_h and $\widetilde{\varphi}_h$ be two embedded RK methods with orders p and q $(1 \le q \le p-1)$.*

(i) *If $\langle y_n, C_{q+1}(y_n)\rangle \neq 0$, then there exists $h^* > 0$ such that the projected method $\widehat{\varphi}_h$ defined by (14) with $\lambda_n = \lambda(y_n, h)$ given by (18) has order $\widehat{p} \ge p$ for all $h \in (0, h^*]$.*

(ii) *If β_n defined by (17) satisfies $\beta_n = B(y_n)h^r + \mathcal{O}(h^{r+1})$ with $B(y_n) \neq 0$, $r \ge 1$, and $p - 2r + 1 > 0$, then the projected method $\widehat{\varphi}_h$ has order $\widehat{p} \ge p - r$ for all $h \in (0, h^*]$.*

Proof. (i) Since $\beta(y_n, 0) \neq 0$ and by (17) we have $\alpha_n = \mathcal{O}(1)$, $\delta_n = \mathcal{O}(h^{p+1})$, there exist some $h^* > 0$ such that $\beta_n^2 - \alpha_n\delta_n > 0$ for $h \in (0, h^*]$. Then λ_n is well defined by (18) and $\lambda_n = \mathcal{O}(h^{p+1})$, which implies that the projected method (14) has order $\widehat{p} \ge p$.

(ii) Now, since $\beta_n = \mathcal{O}(h^r)$, $\alpha_n = \mathcal{O}(1)$, $\delta_n = \mathcal{O}(h^{p+1})$, and $2r < p+1$, $\beta_n^2 - \alpha_n\delta_n > 0$ also holds for $h \in (0, h^*]$. Further it follows from (18) that $\lambda_n = \mathcal{O}(h^{p+1-r})$ and now the order $\widehat{p}$ of the projected method is $\ge p - r$. ∎

In practical application of (14), given an explicit *RK* method φ_h of order p, we usually have the freedom to choose several embedded methods $\widetilde{\varphi}_h$ with different orders $(1 \le q \le p-1)$. Then the question is how to choose $\widetilde{\varphi}_h$. It will be seen that the simplest choice, namely Euler's method, in general provides a projected method with order $\ge p$. In fact, for $\widetilde{\varphi}_h(y_n) = y_n + hf(y_n)$, we have $q = 1$ and $C_2 = y_n''(t_n)/2 = f'(f(y_n))/2$, and the condition

$$\beta(y_n, 0) = \left\langle y_n, \frac{f'(f(y_n))}{\|f'(f(y_n))\|_2} \right\rangle \neq 0$$

implies that the projected method has order $\ge p$.

Finally, observe that $\langle y_n(t+h), y_n(t+h)\rangle - \langle y_n, y_n\rangle = 0$ along the local solution $y_n(t)$ at (t_n, y_n) for all h sufficiently small implies $\langle y_n, f'(f(y_n))\rangle = -\langle f(y_n), f(y_n)\rangle$ and the above condition can be easily checked.

Remark. The above theory extends easily to more general invariants which include both quadratic and linear terms with the form $G(y) = y^T Sy + d^T y$, where S is a symmetric constant matrix and d a constant vector. In fact, for $\widehat{\varphi}_h$ given by (14), (15) holds if and only if λ_n satisfies

$$\alpha_n\lambda_n^2 - 2\beta_n'\lambda_n + \delta_n' = 0, \tag{20}$$

with $\beta_n' = \beta_n + (1/2)d^T w_n$ and $\delta_n' = \delta_n + d^T\varphi_h(y_n) - d^T y_n$. If $\delta_n' = 0$, for a given h, the original method φ_h preserves the invariant and we take $\lambda_n = 0$. Otherwise, the existence of a real solution of the quadratic equation (20) depends on the sign of $(\beta_n')^2 - \alpha_n\delta_n'$, and sufficient conditions similar to those in Theorem 3.1 can be given in order to guarantee the solvability.

4. General invariants. In this section we extend our technique to nonquadratic scalar invariants and to the case of several scalar invariants.

For a general scalar invariant given by a smooth function $G = G(y) : \widehat{\mathcal{D}} \subset \mathbb{R}^m \to \mathbb{R}$ the proposed projection technique reduces to the solution of the nonlinear scalar equation

$$G\Big(\varphi_h(y_n) - \lambda_n w_n\Big) - G(y_n) = 0, \tag{21}$$

where $w_n = w_n(y_n, h)$ is given by (13). In this case, the existence of solution λ_n is guaranteed under the conditions of the following theorem.

THEOREM 4.1. *Let φ_h and $\widetilde{\varphi}_h$ be two embedded RK methods with orders p and q $(1 \le q \le p-1)$, respectively, and let $C_{q+1}(y_n)h^{q+1}$ be the leading term of the local error $\widetilde{\varphi}_h(y_n)$.*

(i) *If $\nabla G(y_n)^T C_{q+1}(y_n) \neq 0$, then there exist $h^* > 0$ such that (21) defines a unique function $\lambda_n = \lambda(y_n, h)$ for all $h \in [0, h^*]$ and the projected method $\widehat{\varphi}(y_n)$ has order $\hat{p} \ge p$.*

(ii) *If $\nabla G(\varphi_h(y_n))^T w_n = B(y_n)h^r + \mathcal{O}(h^{r+1})$ with $r \ge 1$, $B(y_n) \neq 0$, and $2r \le p$, there exist $h^* > 0$ such that (21) defines a unique function $\lambda_n = \lambda(y_n, h)$ for all $h \in (0, h^*]$ and the projected method $\widehat{\varphi}(y_n)$ has order $\hat{p} \ge p - r$.*

Proof. (i) For λ, h in a neighborhood of the origin let us define the real function $g(\lambda, h) \equiv G(\varphi_h(y_n) - \lambda w_n) - G(y_n)$. Taking into account the smoothness of G and φ_h and that

$$g(0,0) = G(y_n) - G(y_n) = 0, \quad \frac{\partial g}{\partial \lambda}(0,0) = -\nabla G(y_n)^T \frac{C_{q+1}(y_n)}{\|C_{q+1}(y_n)\|_2} \neq 0,$$

the implicit function theorem ensures the existence of a neighborhood $[0, h^*]$ and a unique smooth function $\lambda_n = \lambda_n(h)$ satisfying $\lambda_n(0) = 0$ and $g(\lambda_n(h), h) = 0$ for all $h \in [0, h^*]$.

Moreover, we can expand

$$g(\lambda_n, h) = g(0, h) + \frac{\partial g}{\partial \lambda_n}(0, h)\lambda_n + \mathcal{O}(\lambda_n^2),$$

with

$$g(0,h) = G(\varphi_h(y_n)) - G(y_n) = \mathcal{O}(h^{p+1}),$$
$$\frac{\partial g}{\partial \lambda_n}(0,h) = \frac{\partial g}{\partial \lambda_n}(0,0) + \mathcal{O}(h).$$

Then $\lambda_n = \lambda_n(h) = \mathcal{O}(h^{p+1})$ and the projected method has order $\hat{p} \ge p$.

(ii) Now we consider the real function $z(\mu, h)$ defined in the neighborhood of the origin by

$$z(\mu, h) = h^{-2r}\left[G(\varphi_h(y_n) - \mu h^r w_n) - G(y_n)\right] \quad \text{for } h \neq 0, \tag{22}$$

where $z(\mu, 0)$ is given by the corresponding limit

$$\lim_{h \to 0} z(\mu, h) = -B(y_n)\mu + \frac{1}{2\|C_{q+1}(y_n)\|_2^2} C_{q+1}(y_n)^T \frac{\partial^2 G}{\partial y^2}(y_n) C_{q+1}(y_n)\mu^2. \tag{23}$$

Clearly z is continuous for all $h > 0$. To check the continuity of z at $h = 0$ note that by the mean value theorem,

$$G(\varphi_h(y_n) - \mu h^r w_n) = G(\varphi_h(y_n)) - \mu h^r \int_0^1 \nabla G(\varphi_h(y_n) - \theta \mu h^r w_n)^T w_n \mathrm{d}\theta;$$

then for $h \neq 0$ we have

$$z(\mu, h) = -\mu h^{-r} \int_0^1 \nabla G(\varphi_h(y_n) - \theta \mu h^r w_n)^T w_n \mathrm{d}\theta + h^{-2r}\big(G(\varphi_h(y_n)) - G(y_n)\big).$$

Let us consider the smooth function defined as $\varsigma(y) := \nabla G(y)^T w_n$. By considering again the mean value theorem applied to ς, we can write

$$\varsigma(\varphi_h(y_n) - \theta\mu h^r w_n) = \varsigma(\varphi_h(y_n)) - \theta\mu h^r \int_0^1 \nabla\varsigma(\varphi_h(y_n) - \rho\theta\mu h^r w_n)^T w_n \mathrm{d}\rho.$$

Thus, we have that

$$\begin{aligned} z(\mu,h) = {} & h^{-2r}(G(\varphi_h(y_n)) - G(y_n)) - \mu h^{-r}\varsigma(\varphi_h(y_n)) \\ & + \mu^2 \int_0^1 \int_0^1 \theta \nabla\varsigma(\varphi_h(y_n) - \rho\theta\mu h^r w_n)^T w_n \mathrm{d}\rho\mathrm{d}\theta. \end{aligned} \tag{24}$$

Since $p+1 > 2r$, we deduce (23) from (24), and then $z(\mu,h)$ is a continuous function at $h = 0$.

Moreover, $\partial z/\partial\mu$ is also a continuous function in a neighborhood of the origin, and $\partial z/\partial\mu(0,0) = -B(y_n) \neq 0$; therefore by the implicit function theorem there exist $h^* > 0$ and a continuous $\mu = \mu(h)$ such that $z(\mu(h),h) = 0$ for all $h \in [0,h^*]$.

Taking into account (22), for all $h \in (0,h^*]$, $\lambda_n(h)$ given by $\lambda_n(h) = h^r\mu(h)$ is the unique solution of $G(\varphi_h(y_n) - \lambda_n w_n) - G(y_n) = 0$.

Finally, by considering the expansion

$$z(\mu,h) = z(0,h) + \frac{\partial z}{\partial\mu}(0,h)\mu + \mathcal{O}(\mu^2),$$

with

$$\begin{aligned} z(0,h) &= h^{-2r}\big(G(\varphi_h(y_n)) - G(y_n)\big) = \mathcal{O}(h^{p+1-2r}), \\ \frac{\partial z}{\partial\mu}(0,h) &= \frac{\partial z}{\partial\mu}(0,0) + \mathcal{O}(h) = -B(y_n) + \mathcal{O}(h), \end{aligned}$$

it follows that $\mu = \mathcal{O}(h^{p+1-2r})$ and therefore $\lambda_n = \mathcal{O}(h^{p+1-r})$, which implies that the projected method has order $\hat{p} \geq p - r$. ∎

Remarks.

1. Since w_n is a unit vector in the direction of the local error of the embedded method $\widetilde{\varphi}_h(y_n)$ and $w_n = C_{q+1}(y_n)/\|C_{q+1}(y_n)\|_2 + \mathcal{O}(h)$, $C_{q+1}(y_n) \neq 0$, condition $\nabla G(y_n)^T C_{q+1}(y_n) \neq 0$ of Theorem 4.1(i) is equivalent to the nonvanishing of the derivative of $G(y)$ in the direction of the leading error term $C_{q+1}(y_n)$ of w_n at y_n. Geometrically it means that $C_{q+1}(y_n)$ is not tangent to the level surface $G(y) - G(y_n) = 0$ at y_n. Further, if the leading term of $\nabla G(\varphi_h(y_n))^T w_n$ given by $\nabla G(y_n)^T C_{q+1}(y_n)$ vanishes for some y_n, but $\nabla G(\varphi_h(y_n))^T w_n = \mathcal{O}(h^r)$ with $1 \leq r \leq p/2$, it is possible to ensure the preservation of the invariant at the cost of an order reduction ($\hat{p} \geq p - r$) of the projected numerical solution.
2. Theorem 4.1 can also be applied to prove that the standard orthogonal projection of an RK method of order p has order $\geq p$. In fact (5) has the form (14) with $w_n = -\nabla G(\varphi_h(y_n))$, and in this case since $\nabla G(y_n)^T w_n = \|\nabla G(y_n)\|_2^2 + \mathcal{O}(h)$, the method (5) has order $\geq p$. If $\nabla G(y_n) = 0$, the standard projection cannot be applied, but in such a case if $\nabla G(\varphi_h(y_n))^T w_n = \mathcal{O}(h^r)$, our projection technique could be applied. Note also that the standard orthogonal projection requires having an analytical expression of $\nabla G(y)$, whereas in our approach the gradient is not necessary.

3. Recall that for all affine transformations $z \to y = Pz + q$ with constant $P \in \mathbb{R}^{m\times m}$, $q \in \mathbb{R}^m$, if we denote by $\psi_{f,t}$ the flow map of $y' = f(y)$ and by $\psi_{\tilde{f},t}$ the corresponding flow map to the transformed system $z' = \tilde{f}(z) = P^{-1}f(Pz+q)$, they are affine invariants in the sense that $\psi_{f,t}(y_0) = P\psi_{\tilde{f},t}(z_0) + q$ whenever $y_0 = Pz_0+q$. It can be seen that this property holds for RK methods and also for the proposed methods (14). However, the standard projection preserves the affine invariance only for orthogonal P.

4.1. Implementation in adaptive RK codes. To use our technique in adaptive codes based on an embedded pair of RK formulas with orders p and $p-1$, recall that the step size $t_n \to t_{n+1} = t_n + h_n$ is chosen so that a norm of the local error estimator $\mathrm{EST} = y_{n+1} - y_{n+1}^{(p-1)}$ given by the difference between the solutions y_{n+1} and $y_{n+1}^{(p-1)}$ of orders p and $p-1$, respectively, is lower than a given error tolerance TOL. For the new projected solution $\widehat{y}_{n+1} = y_{n+1} - \lambda_n w_n$ one is tempted to use the new estimator $\widehat{\mathrm{EST}} = \widehat{y}_{n+1} - y_{n+1}^{(p-1)}$, but $\widehat{y}_{n+1}$ could lose (at some isolated times) one or more orders of accuracy and then this is not a reliable estimate. Thus we propose a more conservative policy in which we check both EST and the correction of projection $\|\lambda_n w_n\| = |\lambda_n|$ and accept a step whenever $\max\{\|\mathrm{EST}\|, |\lambda_n|\} \le \mathrm{TOL}/2$.

Observe that $\|\mathrm{EST}\| = \mathcal{O}(h^p)$ and that if the projected solution has no order reduction, then $|\lambda_n| = \mathcal{O}(h^{p+1})$. Hence the above step-size control will be usually governed by EST.

Then for a rejected step h_n' or the new step h_{n+1} we propose the new step size given by the modified standard formula

$$\mathrm{fac} * \left(\frac{\mathrm{TOL}}{2\max\{\|\mathrm{EST}\|, \lambda_n\}}\right)^{1/p} * h_n,$$

with the security factor fac of the original code.

On the other hand, note that in the projected method $\widehat{y}_{n+1} = \widehat{\varphi}_h(y_n)$ can be rewritten as

$$\widehat{y}_{n+1} = y_{n+1} - \lambda_n w_n = y_{n+1} - \frac{\lambda_n}{\|y_{n+1} - \widetilde{y}_{n+1}\|}\,(y_{n+1} - \widetilde{y}_{n+1})$$

and is an RK type method whenever $\lambda_n \|y_{n+1} - \widetilde{y}_{n+1}\|^{-1}$ is bounded. Since $\lambda_n = \mathcal{O}(h^{p+1-r})$ and $\|y_{n+1} - \widetilde{y}_{n+1}\| = \mathcal{O}(h^{q+1})$, the above quotient is $\mathcal{O}(h^{p-q-r})$; then for a strong order reduction ($r > p-q$) the projected method would lose its reliability. To protect the code from such eventual strong order reductions we propose to include the additional control $|\lambda_n| < \|y_{n+1} - \widetilde{y}_{n+1}\|$. If this condition does not hold at some step, we would take another $\widetilde{\varphi}_h$ method that would change the direction of the projected method. In our numerical experiments we have not detected such a strong order reduction that this control has been necessary.

Finally, in practical application of our technique with one scalar nonquadratic invariant we must solve at each step the (nonquadratic) equation $g(\lambda, h) = 0$. Many iterative methods have been proposed to solve nonlinear equations: In Newton-type methods we need to approximate the gradient of G (either analytically or numerically) one or more times per step depending on the convergence of the iteration. In our numerical experiments with the modified Newton, convergence was attained with 1–2 iterations. However, other iterative schemes that require only evaluations of G may be used. We have found that the secant method, with starting values $\lambda = 0$ and the

value λ obtained in the previous step, requires 1–2 iterations to attain convergence to the round-off level in the tested problems. Thus, we have used this iterative scheme because its convergence is similar to that of the modified Newton and it has a lower computational cost.

4.2. Several invariants. Suppose now that the differential system (1) possesses $l \geq 1$ smooth invariants $G_1(y), \ldots, G_l(y)$ defined in $\mathcal{D} \subset \mathbb{R}^m$, and let $G = (G_1, \ldots, G_l)^T : \mathcal{D} \subset \mathbb{R}^m \to \mathbb{R}^l$. Together with the starting method φ_h of order p we consider l linearly independent embedded methods $\widetilde{\varphi}_h^{(1)}, \ldots, \widetilde{\varphi}_h^{(l)}$ with orders $q_1 \leq q_2 \leq \cdots \leq q_l < p$. Then, generalizing (14), the projected method $\widehat{\varphi}_h$ will be defined by

$$\widehat{\varphi}_h(y_n) = \varphi_h(y_n) - \sum_{i=1}^{l} \lambda_n^{(i)} w_n^{(i)}, \tag{25}$$

where

$$w_n^{(i)} = \frac{\varphi_h(y_n) - \widetilde{\varphi}_h^{(i)}(y_n)}{\|\varphi_h(y_n) - \widetilde{\varphi}_h^{(i)}(y_n)\|}$$

and $\lambda_n^{(i)} = \lambda^{(i)}(y_n, h)$, $i = 1, \ldots, l$, are problem-dependent scalars that will be chosen so that

$$G\left(\varphi_h(y_n) - \sum_{i=1}^{l} \lambda_n^{(i)} w_n^{(i)}\right) = G(y_n). \tag{26}$$

We also introduce the $m \times l$ matrix T_n defined by its column vectors $T_n = [w_n^{(1)} | \ldots | w_n^{(l)}]$, and then

$$\left(\nabla G(\varphi_h(y_n))\right)^T \; T_n = M(y_n) + \mathcal{O}(h), \tag{27}$$

with $M(y_n) \in \mathbb{R}^{l \times l}$. With this notation we have the following theorem.

THEOREM 4.2. *If* (27) *holds with nonsingular* $M(y_n)$*, then there exist* $h^* > 0$ *and a unique set* $(\lambda^{(1)}(y_n, h), \ldots, \lambda^{(l)}(y_n, h))$ *such that the projected method* $\widehat{\varphi}_h$ *defined by* (25) *preserves the* l *invariants* $G_i, i = 1, \ldots, l$*, for all* $h \in [0, h^*]$ *and has order* $\geq p$*.*

Proof. Consider $\lambda = (\lambda^{(1)}, \ldots, \lambda^{(l)})^T$ and $g(\lambda, h) : \mathbb{R}^l \times \mathbb{R} \to \mathbb{R}$ defined by

$$g(\lambda, h) = G\left(\varphi_h(y_n) - \sum_{i=1}^{l} \lambda^{(i)} w_n^{(i)}\right) - G(y_n). \tag{28}$$

The proof follows in a way similar to that of statement (i) of Theorem 4.1. ∎

To end this section, the following result is given in order to ensure that Newton-type iterations applied to (26) provide numerical sequences which converge in small neighborhoods of the origin. In fact, we have the following corollary.

COROLLARY 4.1. *Under the same conditions of Theorem* 4.2 *above, there exists* $h^* > 0$ *such that for all* $0 < h < h^*$ *the sequence defined by the Newton method applied to* (26)*, with starting value* $\lambda := 0$*, converges to the unique solution of this equation in a sufficiently small neighborhood of the origin.*

Proof. Considering, as in Theorem 4.2 above, the function $g(\lambda, h)$ (28), it is not difficult to see that

$$\gamma := \sup_{\substack{\eta,\mu\in\Lambda_0 \\ \eta\neq\mu}} \frac{\| \nabla_\lambda g(\eta, h) - \nabla_\lambda g(\mu, h) \|}{\| \eta - \mu \|} = \mathcal{O}(1),$$

$$\alpha := \| \nabla_\lambda g(0, h)^{-1} g(0, h) \| = \mathcal{O}(h^{p+1}),$$

$$\beta := \| \nabla_\lambda g(0, h)^{-1} \| = \mathcal{O}(1),$$

where $\Lambda_0 \subset \mathbb{R}^l$ stands for a sufficiently small neighborhood of $\lambda = 0$. Thus, there exists $h^* > 0$ such that for all $0 < h < h^*$, $\tau := \gamma\alpha\beta = \mathcal{O}(h^{p+1}) \leq 1/2$ is fulfilled. Then, the statement follows from the Newton–Kantorovich theorem (see, e.g., Theorem 5.3.6 in [14]). ∎

Concerning the implementation of the new projection technique with several invariants, it is straightforward to extend the step-size control of subsection 4.1 to several invariants.

5. Numerical experiments. Here we present some numerical experiments with the purpose of showing some features of the new projection technique. First of all we want to make clear that it is possible to include in any adaptive RK code the new projection technique in a simple way. Further, the resulting code retains its reliability and accuracy and has a low additional computational cost. For comparison we have also included the standard orthogonal projection technique which also preserves the order of the original method but requires explicit expressions of the gradients ∇G and in general has a higher computational cost. In our experiments we have employed as adaptive RK code the pair of Dormand and Prince, DoPri5(4) (see, e.g., [4] for details on the coefficients of this method). Some practical examples have been considered in order to illustrate how the projection technique performs on this kind of problem with invariants. For the sake of brevity, we will show here some results related to the following three problems. Further numerical examples can be found in [2].

Problem 1. A problem of micromagnetism [12].

$$f(y) := H_{eff} \times y + \lambda y \times (H_{eff} \times y),$$

with $\lambda = 1/20.1$ and $H_{eff}, y \in \mathbb{R}^3$. We have chosen the initial value $y(0) = (y_0[1], y_0[2], y_0[3]) = (\sin\theta_0 \cos\varphi_0, -\sin\theta_0 \sin\varphi_0, \cos\theta_0)^T$, where $\varphi_0 = \pi/4$ and $\theta_0 = \pi/3$.

This kind of equation appears in micromagnetism problems when solving the Landau–Lifshitz–Gilbert equation (see, e.g., [8, 12, 13]). Since $y'(t)$ is orthogonal to the exact solution $y(t)$, it is clear that $G(y) = y^T y$ is a quadratic invariant. In case of $H_{eff} := (1, 0, 0)^T$, the exact solution is given by

$$y(t) = \left(\frac{a(t)}{b(t)}, \frac{2}{b(t)}(y_0[2]\cos t - y_0[3]\sin t), \frac{2}{b(t)}(y_0[2]\sin t + y_0[3]\cos t) \right)^T,$$

with

$$a(t) = e^{\lambda t}(1 + y_0[1]) - e^{-\lambda t}(1 - y_0[1]),$$
$$b(t) = e^{\lambda t}(1 + y_0[1]) + e^{-\lambda t}(1 - y_0[1]).$$

Problem 2. The restricted three body problem (see, e.g., [6, pp. 129–130]).
Here $F : \mathbb{R}^4 \to \mathbb{R}^4$ is given by

$$f(y) := \left(y_3, y_4, y_1 + 2y_4 - \bar{\mu}\frac{y_1+\mu}{D_1^{3/2}} - \mu\frac{y_1-\bar{\mu}}{D_2^{3/2}}, y_2 - 2y_3 - \bar{\mu}\frac{y_2}{D_1^{3/2}} - \mu\frac{y_2}{D_2^{3/2}} \right)^T,$$

with $D_1 := (y_1+\mu)^2 + y_2^2$, $D_2 := (y_1-\bar{\mu})^2 + y_2^2$, $y = (y_1, y_2, y_3, y_4)^T$, and $y(0) = (0.994, 0, 0, -2.00158510637908252240537862224)^T$, where $\bar{\mu} = 1-\mu$ and $\mu = 0.012277471$. For this initial condition we get a periodic solution with a minimal period $T_2 := 17.06521656015796255889172062 49$. Now the total energy

$$G(y) = \frac{1}{2}(y_3^2 + y_4^2 - y_1^2 - y_2^2) - \bar{\mu}D_1^{-1/2} - \mu D_2^{-1/2}$$

is a nonquadratic first integral.

Problem 3. Euler equations (see, e.g., [1] and [5, pp. 95–96]).

$$f(y) := \big((\alpha-\beta)y_2y_3,\ (1-\alpha)y_3y_1,\ (\beta-1)y_1y_2\big)^T,$$

with $y = (y_1, y_2, y_3)^T$ and $y(0) = (0, 1, 1)^T$, where $\alpha = 1 + \frac{1}{\sqrt{1.51}}$ and $\beta = 1 - \frac{0.51}{\sqrt{1.51}}$. In these equations the exact solution can be written as

$$y(t) = \left(\sqrt{1.51}\ \mathrm{sn}(t, 0.51), \mathrm{cn}(t, 0.51), \mathrm{dn}(t, 0.51) \right)^T,$$

where sn, cn, dn stand for the elliptic Jacobi functions. This problem has two quadratic integrals:

$$\text{(i)}\ \ G_1(y) = y_1^2 + y_2^2 + y_3^2, \qquad \text{(ii)}\ \ G_2(y) = y_1^2 + \beta y_2^2 + \alpha y_3^2.$$

These problems have been integrated with the following three methods: DoPri5(4) (the classical pair of Dormand and Prince [4]), DoPri5(4)+SP (the pair of Dormand and Prince with the standard projection technique), and DoPri5(4)+DP (the same pair with the proposed directional projection technique). To define the directional projection we have used the Euler method as embedded method $\widetilde{y}_{n+1}$ (of order $q = 1$) when only one scalar invariant must be preserved (Problems 1 and 2). For Problem 3, which posses two invariants, we need two "independent" embedded methods of orders q_1 and q_2 to define the projected method. We have used the Euler method ($q_1 = 1$) and a second-order embedded method ($q_2 = 2$).

For a quadratic invariant $G(y) = y^T Sy$, since the gradient $\nabla G = 2Sy$ is immediately available, we compute λ_n from the quadratic equation (12). For nonquadratic invariants λ_n is computed by a modified Newton iteration in the case of orthogonal projection and by a secant iterative scheme for our projected solution. For several invariants we have used modified Newton for both projection techniques (in the case of our projected solution Broyden's iteration is under consideration).

The first problem has been integrated with several fixed step sizes in the interval $[0, T_{end} = 16\pi]$. The global and invariant errors (i.e., *ge* and *ie*, resp.) for DoPri5(4) and DoPri5(4)+DP and the different step sizes have been plotted in Figure 1. It can be observed that the projected method preserves the invariant and also provides a numerical solution with the same order (the slopes of the global error curves are the same) but with a smaller global error than classical Dopri5(4).

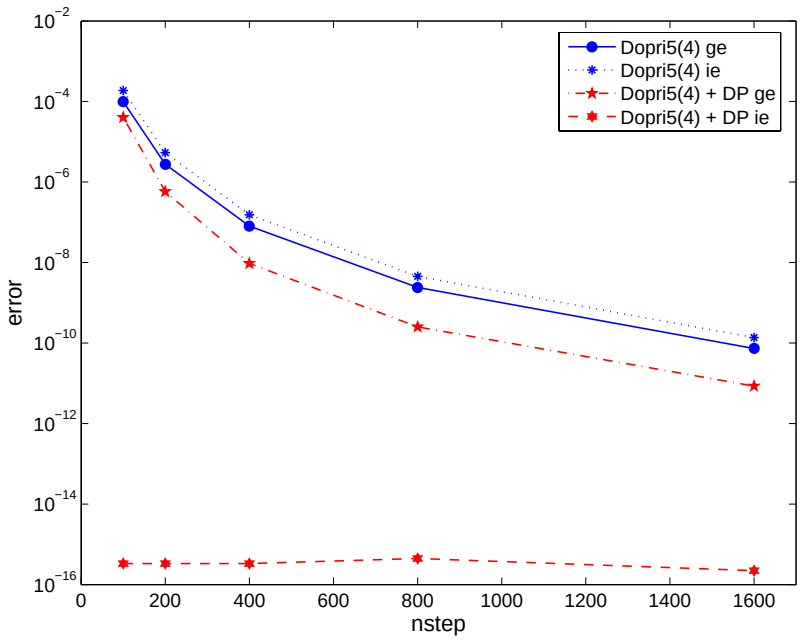

FIG. 1. *Problem* 1 *integrated with DoPri*5(4) *and DoPri*5(4)+*DP.*

On the other hand, Problems 2 and 3 have been integrated with variable step size using the technique proposed in subsection 4.1, with values $ATOL = 10^{-6}$ and $RTOL = ATOL/10$ representing the absolute and relative tolerance errors, respectively. The fourth-order embedded method was considered in order to estimate the local error at each point in the nonuniform grid. The end point was taken in each problem as $T_{end} = 3T_2$ and $T_{end} = 100$, respectively.

Figures 2 and 3 show the numerical orbits in the (y_1, y_2)-plane when Problem 2 is integrated with DoPri5(4) and DoPri5(4)+DP, respectively (the figure for DoPri5(4)+SP is very similar to the one corresponding to DoPri5(4)+DP). It is seen that both projected methods give a better qualitative behavior than DoPri5(4) because they preserve the corresponding first integral, as shown in Figure 4. It must be remarked that some round-off errors came up when solving the nonlinear equations defining the projected methods in neighborhoods of the singularities of the underlying problem, and this is the reason the projected methods do not preserve properly the invariant in neighborhoods of the singularities. Efficiency curves have not been presented for this problem because they show a similar behavior to those in Problems 1 and 3.

Figures 5 and 6 correspond to Problem 3, integrated with variable step size. In this case, we have displayed in Figure 5 the global error produced by the three methods in relation to the number of function evaluations, whereas Figure 6 plots the error in both invariants. Again, the projected methods provide numerical solutions with a smaller global error than the classical pair 5(4) for a similar number of derivative function evaluations.

6. Conclusions. A projection technique to preserve invariants by means of explicit RK methods has been proposed. The projected methods preserve both the order

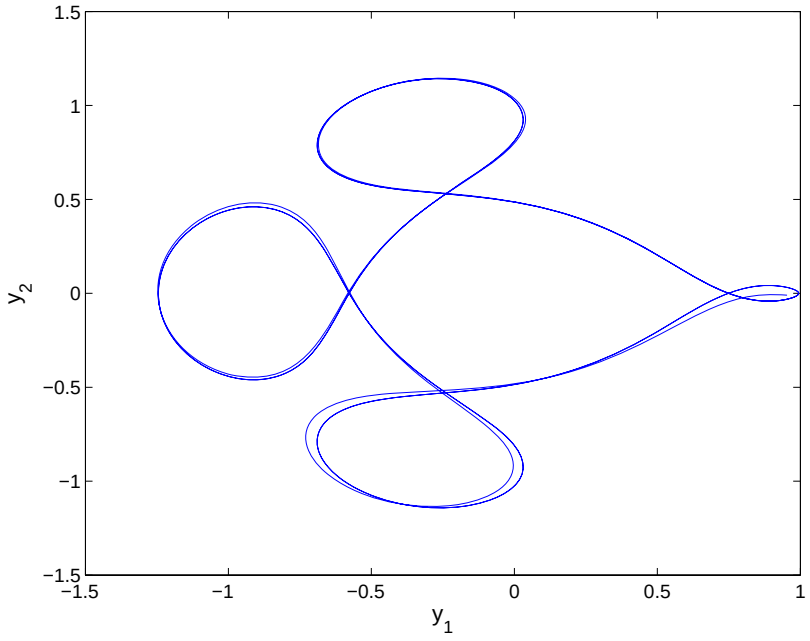

FIG. 2. *Problem* 2 *integrated with DoPri*5(4).

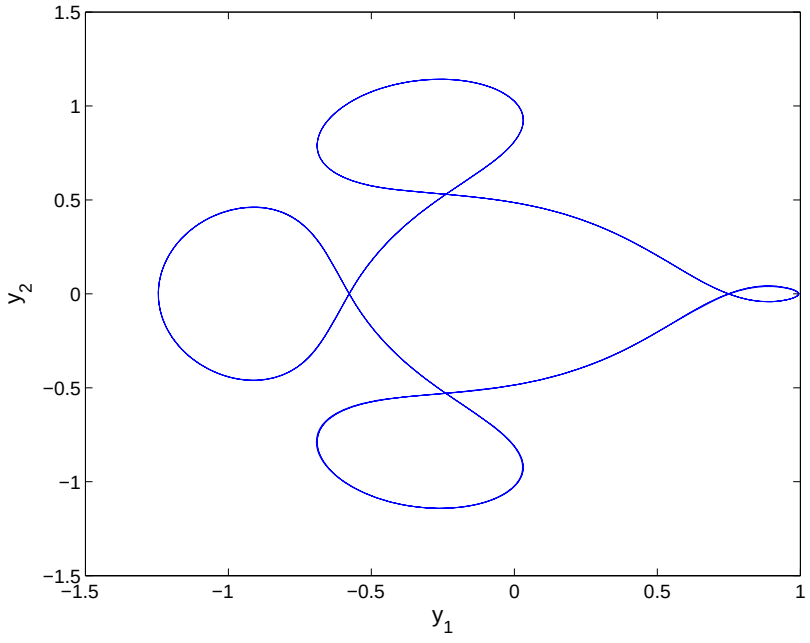

FIG. 3. *Problem* 2 *integrated with DoPri*5(4)+*SP or DoPri*5(4)+*DP*.

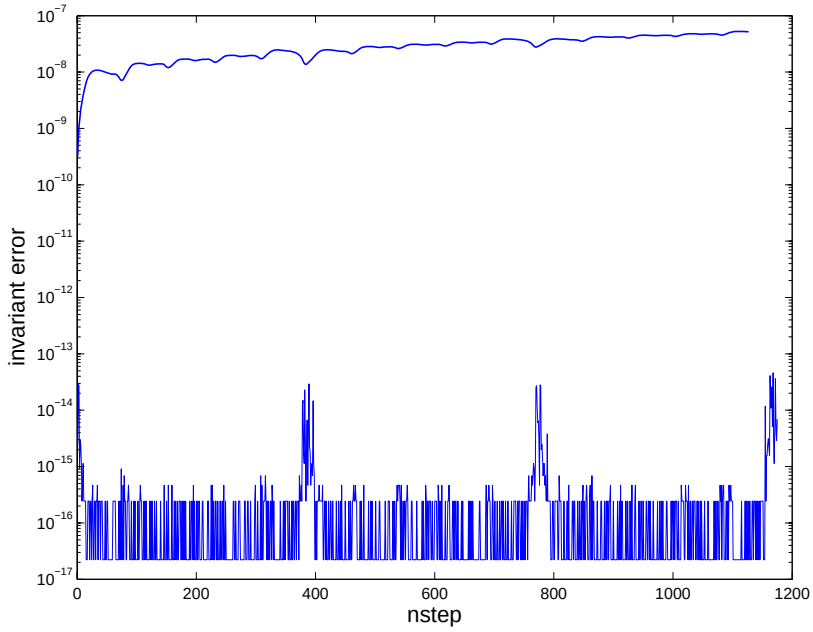

FIG. 4. *Problem* 2. *Invariant errors with DoPri*5(4) *and DoPri*5(4)+*DP.*

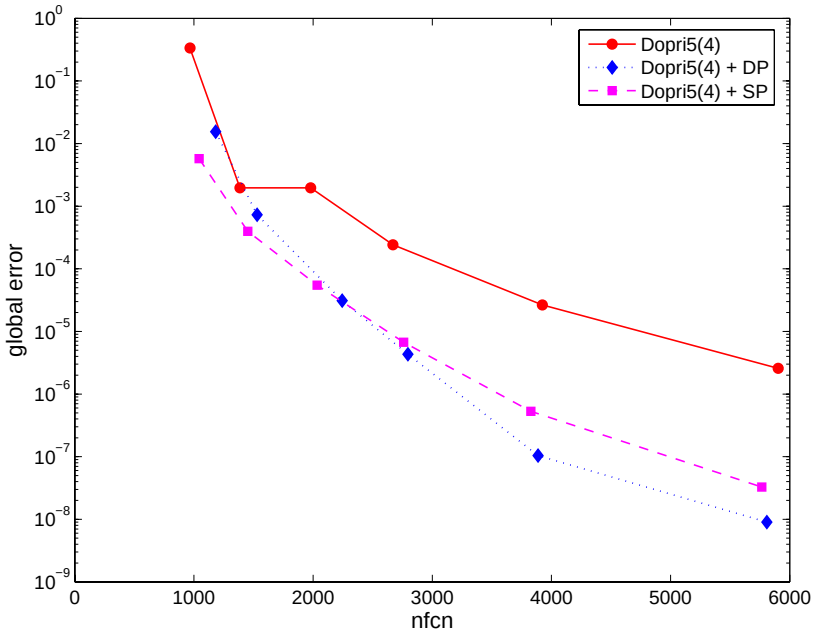

FIG. 5. *Problem* 3. *Global errors with DoPri*5(4), *DoPri*5(4)+*SP, and DoPri*5(4)+*DP.*

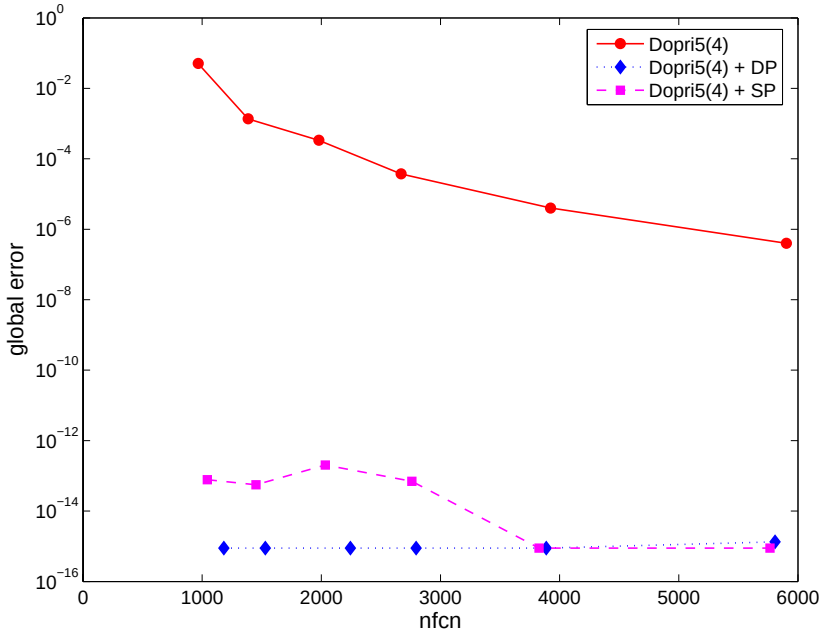

FIG. 6. *Problem* 4. *Invariant errors with DoPri*5(4), *DoPri*5(4)+*SP, and DoPri*5(4)+*DP.*

of consistency of the original methods and the selected invariants of the problem under consideration. A study of the order of consistency has been carried out.

The main advantages of the new approach over the standard projection technique are as follows: The gradient of the invariants is not required. The approach preserves linear invariants and affine invariance. Further, the new projection methods are well suited for use in adaptive RK methods with a low extra cost.

Finally the new technique and the standard projection have been implemented in the code DoPri5(4), and the results of some numerical experiments are presented.

REFERENCES

[1] N. DEL BUONO AND C. MASTROSERIO, *Explicit methods based on a class of four stage fourth order Runge–Kutta methods for preserving quadratic laws*, J. Comput. Appl. Math., 140 (2002), pp. 231–243.

[2] M. CALVO, D. HERNÁNDEZ-ABREU, J. I. MONTIJANO, AND L. RÁNDEZ, *Explicit Runge-Kutta methods for the preservation of invariants*, Technical report, Departamento Matemática Aplicada, Universidad Zaragoza, Zaragoza, Spain, 2004.

[3] G. J. COOPER, *Stability of Runge–Kutta methods for trajectory problems*, IMA J. Numer. Anal., 7 (1987), pp. 1–13.

[4] J. R. DORMAND AND P. J. PRINCE, *A family of embedded Runge-Kutta formulae*, J. Comp. Appl. Math., 6 (1980), pp. 19-26.

[5] E. HAIRER, C. LUBICH, AND G. WANNER, *Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations*, Springer-Verlag, Berlin, 2002.

[6] E. HAIRER, S. P. NØRSETT, AND G. WANNER, *Solving Ordinary Differential Equations*. I. *Non-stiff Problems*, Springer Ser. Comput. Math. 8, Springer-Verlag, Berlin, 1993.

[7] A. ISERLES AND A. ZANNA, *Preserving algebraic invariants with Runge-Kutta methods*, J. Comput. Appl. Math., 125 (2000), pp. 69–81.

[8] D. LEWIS AND N. NIGAM, *Geometric integration on spheres and some interesting applications*, J. Comput. Appl. Math., 151 (2003), pp. 141–170.

[9] G. R. W. Quispel and D. I. McLaren, *Integral-preserving integrators*, J. Phys. A, 37 (2004), pp. L489–L495.
[10] J. Schropp, *Conserving first integrals under discretization with variable step size integration procedures*, J. Comput. Appl. Math., 115 (2000), pp. 503–517.
[11] L. F. Shampine, *Conservation laws and the numerical solution of ODEs*, Comput. Math. Appl. Part B, 12 (1986), pp. 1287–1296.
[12] M. Slodička and I. Cimrák, *An iterative approximation scheme for the Landau-Lifshitz-Gilbert equation*, J. Comput. Appl. Math., 169 (2004), pp. 17–32.
[13] M. Slodička and I. Cimrák, *Numerical study of nonlinear ferromagnetic materials*, Appl. Numer. Math., 46 (2003), pp. 95–111.
[14] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1983.
[15] A. Wintner, *The Analytical Foundations of Celestial Mechanics*, Princeton University Press, Princeton, NJ, 1941.