Found Comput Math (2008) 8: 335–355
DOI 10.1007/s10208-007-9009-6

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Explicit Volume-Preserving Splitting Methods for Linear and Quadratic Divergence-Free Vector Fields

R.I. McLachlan · H.Z. Munthe-Kaas · G.R.W. Quispel · A. Zanna

Received: 24 November 2006 / Revised: 6 August 2007 / Accepted: 14 August 2007 /
Published online: 22 November 2007

Abstract We present new explicit volume-preserving methods based on splitting for polynomial divergence-free vector fields. The methods can be divided in two classes: methods that distinguish between the diagonal part and the off-diagonal part and methods that do not. For the methods in the first class it is possible to combine different treatments of the diagonal and off-diagonal parts, giving rise to a number of possible combinations.

Keywords Geometric integration · Volume preservation · Splitting methods

AMS Subject Classification 65L05 · 34C14

This paper is dedicated to Arieh Iserles on the occasion of his 60th anniversary.

Communicated by Peter Olver.

R.I. McLachlan
Institute of Fundamental Sciences, Massey University, Palmeston North, New Zealand
e-mail: r.mclachlan@massey.ac.nz

H.Z. Munthe-Kaas · A. Zanna (✉)
Matematisk Institutt, J. Brunsgt 12, 5008 Bergen, Norway
e-mail: anto@math.uib.no

H.Z. Munthe-Kaas
e-mail: hans@math.uib.no

G.R.W. Quispel
Department of Mathematical and Statistical Sciences, La Trobe University, Victoria 3086, Australia
e-mail: R.Quispel@latrobe.edu.au

1 Introduction

In recent years, a new branch of the field of numerical solution of differential equations has come into existence. This branch is called 'geometric integration'. This term denotes the numerical solution of a differential equation or class of differential equations while preserving one (or more) of their properties exactly (i.e. to machine accuracy). Some important examples are symplectic integrators (preserving symplectic structure exactly), integral-preserving integrators (preserving first integrals such as energy and/or (angular) momentum exactly), and volume-preserving integrators (preserving phase-space volume exactly). Recent reviews of geometric integration are [1, 5, 9, 12]. The most common methods used in geometric integration are the so-called splitting methods [10]. These have the advantage that (as is the case in the present paper) they are usually explicit. Furthermore, recent results indicate that methods based on analytic approximations and, in particular, B-series, cannot be volume-preserving [3, 6].

In this paper we study splitting methods for polynomial divergence-free vector fields. Divergence-free vector fields occur naturally in incompressible fluid dynamics, and preservation of phase-space volume is also a crucial ingredient in many if not all ergodic theorems. An earlier paper on splitting polynomial vector fields is [11]. That paper had some discussion of the divergence-free case, but mainly dealt with the Hamiltonian case. Implicit volume-preserving integrators for general divergence-free vector fields were given in [8, 13, 16].

Investigations of the Hamiltonian case, which involves expressing a scalar polynomial of degree d in n variables as a sum of functions of fewer variables, have shown that good splitting methods exist, but that finding and analyzing them (especially for general n and d) is very difficult [4, 11, 14]. The volume-preserving case, which involves n polynomials subject to the divergence-free condition, is even harder, although there is a conjecture [11] that they can be expressed as a sum of $n + d$ shears, each a function of $n - 1$ variables. Therefore, in the present paper we undertake a study of several possible approaches to splitting linear and quadratic divergence-free vector fields in any dimension n, with a view to understanding the structure of the problem and identifying promising approaches for the general case.

We will present several explicit methods that can be divided in two classes: (a) methods that distinguish the diagonal and off-diagonal parts; and (b) methods that do not. By diagonal part we mean all the terms of the vector field such that $\dot{x}_i$ depends on x_i for $i = 1, \ldots, n$. Similarly, the off-diagonal part refers to all the terms of the vector field such that $\dot{x}_i$ does not depend on x_i, $i = 1, \ldots, n$. As far as class (a) is concerned, we introduce several explicit schemes that treat the diagonal part and the off-diagonal part separately. This means that one can pick and choose a method for the diagonal part and then a method for the off-diagonal part, giving rise to a combinatorial number of new methods.

2 Linear Volume-Preserving Vector Fields

In this section we consider several explicit, volume-preserving splitting methods for solving the linear, divergence-free ordinary differential equation (ODE),

$$\dot{\mathbf{x}} = A\mathbf{x}, \quad \operatorname{tr} A = 0. \tag{1}$$

This section is organized as follows. We first analyse methods to treat the off-diagonal part of A and thereafter methods to treat the diagonal part of A (methods of class (a)), Sects. 2.1–2.5). In Sect. 2.6 we consider a method of class (b).

2.1 Splitting in Strictly Triangular Systems

To start with, let us consider the case when the matrix A has zero diagonal elements, $a_{i,i} = 0, i = 1, \ldots, n$. A simple volume-preserving method for (1) can be obtained by decomposing the matrix A as

$$A = A_1 + A_2, \tag{2}$$

where A_1 and A_2 are strictly lower and upper triangular matrices, respectively. Clearly, each of the systems

$$\dot{\mathbf{x}} = A_i\mathbf{x}, \quad i = 1, 2,$$

is volume-preserving. Moreover, these simple systems can be approximated and retain volume by any Runge–Kutta (RK) method, as proved below.

Proposition 1 *Assume that the matrix A in (1) is strictly (upper/lower) triangular. Then, any Runge–Kutta (RK) method applied to (1) is volume-preserving. In particular, both Forward Euler and Backward Euler methods are volume-preserving for (1).*

Proof For Forward Euler (FE) we have $\mathbf{x}_1 = \mathbf{x}_0 + hA\mathbf{x}_0$, and the Jacobian of the numerical method is

$$J_{FE} = \frac{\partial \mathbf{x}_1}{\partial \mathbf{x}_0} = I + hA,$$

with determinant equal to one, as A is strictly triangular. Similarly, for Backward Euler (BE), $\mathbf{x}_1 = \mathbf{x}_0 + hA\mathbf{x}_1$ and

$$J_{BE} = \frac{\partial \mathbf{x}_1}{\partial \mathbf{x}_0} = (I - hA)^{-1}$$

which also has determinant equal to one.

Any other RK method applied to (1) reduces to

$$\mathbf{x}_1 = R(hA)\mathbf{x}_0,$$

where $R(z)$ is the stability function of the RK scheme. This stability function is of the form $R(z) = P(z)/Q(z)$, where $P(z) = \det(I - zA_{RK} + z\mathbf{1}\mathbf{b}_{RK}^{\mathrm{T}})$ and $Q(z) =$

$\det(I - zA_{RK})$ are polynomials such that $P(0) = Q(0) = 1$ and $A_{RK}, \mathbf{b}_{RK}$ are the coefficients and the weights of the RK scheme. Now, $\det \partial \mathbf{x}_1 / \partial \mathbf{x}_0 = \det(R(hA)) = \det(P(hA))/\det(Q(hA)) = 1$ as $P(hA)$ and $Q(hA)$ are triangular matrices with ones along the diagonal (powers of strictly triangular matrices are strictly triangular matrices). □

It is important to stress that, for strictly triangular systems, the Backward Euler method is also explicit: for instance, let us focus on the strictly lower triangular system $\dot{\mathbf{x}} = A_1\mathbf{x}$. If $\mathbf{x}_1 = [x_{1;1}, \ldots, x_{1;n}]^{\mathrm{T}}$, we see that the component $x_{1;k}$ depends only on $x_1, \ldots, x_{k-1}$, hence $\mathbf{x}_1$ can be obtained by a sequence of Forward Euler steps,

$$x_{1;k} = x_{0;k} + hf_k(x_{1;1}, \ldots, x_{1;k-1}), \quad k = 1, 2, \ldots, n. \tag{3}$$

Similarly, for the strictly upper triangular system $\dot{\mathbf{x}} = A_2\mathbf{x}$, we find explicitly all the components of $\mathbf{x}_1$ by solving with a step of Forward Euler from the last component $x_{1;n}$ backwards.

2.2 Splitting the Off-Diagonal Using Canonical Directions

Another simple, explicit, volume-preserving method for matrices with zero diagonal elements can be obtained as follows. Write

$$A = R_1 + R_2 + \cdots + R_n,$$

where R_k is the (traceless) matrix whose only nonzero elements are those in row k, which coincides with those in the kth row of A. The system $\dot{\mathbf{x}} = R_k\mathbf{x}$ reduces to

$$\begin{aligned} &\dot{x}_i = 0, \quad i \neq k, \\ &\dot{x}_k = a_{k,1}x_1 + \cdots + a_{k,k-1}x_{k-1} + a_{k,k+1}x_{k+1} + \cdots + a_{k,n}x_n, \end{aligned} \tag{4}$$

which can also be solved exactly and explicitly by a single step of the Forward Euler method.

This method is equivalent to evolving along one direction at a time. The directions chosen are the n canonical directions $\mathbf{e}_i$ in $\mathbb{R}^n$ (canonical shears).

2.3 Splitting by Generalized Polar Decomposition

The method we describe in this subsection was introduced in [17] for the approximation of the matrix exponential of traceless matrices. Introduce the matrix S_k which coincides with the identity matrix, except for the (k,k)-component, which equals -1. Then the matrix $\frac{1}{2}(A - S_kAS_k)$ picks up exactly the kth row and kth column of A, except for the diagonal element, and is zero elsewhere. If the matrix A has zero diagonal elements, we obtain

$$A = P_1 + P_2 + \cdots + P_{n-1}, \tag{5}$$

where the matrices P_k are constructed in the following iterative way:

$$A^{[0]} = A, \qquad P_1 = \frac{1}{2}\big(A^{[0]} - S_1 A^{[0]} S_1\big),$$
$$A^{[k]} = A^{[k-1]} - P_k, \quad P_{k+1} = \frac{1}{2}\big(A^{[k]} - S_k A^{[k]} S_k\big), \quad k = 1, 2, \ldots, n-2.$$

The remainder $A^{[n-1]} = A^{[n-2]} - P_{n-1}$ is the diagonal part of A which, in our case, is the zero matrix. Each system

$$\dot{\mathbf{x}} = P_i \mathbf{x}, \quad i = 1, 2, \ldots, n-1,$$

is divergence-free as P_i is traceless and can be solved exactly by computing the exact exponential of P_i by an Euler–Rodrigues-type formula,

$$\exp(P_j) = \begin{cases} I + \frac{\sinh\alpha_j}{\alpha_j} P_j + \frac{1}{2}\big(\frac{\sinh(\alpha_j/2)}{\alpha_j/2}\big)^2 P_j^2, & \text{if } \mu_j > 0,\ \alpha_j = \sqrt{\mu_j}, \\ I + P_j + \frac{1}{2} P_j^2, & \text{if } \mu_j = 0, \\ I + \frac{\sin\alpha_j}{\alpha_j} P_j + \frac{1}{2}\big(\frac{\sin(\alpha_j/2)}{\alpha_j/2}\big)^2 P_j^2, & \text{if } \mu_j < 0,\ \alpha_j = \sqrt{-\mu_j}, \end{cases} \tag{6}$$

where $\mu_j = \sum_{k=j+1}^{n} a_{j,k} a_{k,j}$.

Here we will give another interpretation of the method. Consider, for instance, the system $\dot{\mathbf{x}} = P_1 \mathbf{x}$. This reads

$$\begin{aligned} \dot{x}_1 &= a_{1,2} x_2 + \cdots + a_{1,n} x_n, \\ \dot{x}_2 &= a_{2,1} x_1, \\ &\vdots \\ \dot{x}_n &= a_{n,1} x_1. \end{aligned}$$

Differentiating the first equation and substituting the values of $\dot{x}_i$, $i = 2, \ldots, n$, yields the linear, second-order differential equation for x_1,

$$\ddot{x}_1 - \mu_1 x_1 = 0, \quad \mu_1 = \sum_{i=2}^{n} a_{1,i} a_{i,1},$$

which has solution $x_1(t) = c_1 \mathrm{e}^{\lambda_1 t} + c_2 \mathrm{e}^{\lambda_2 t}$, where λ_1 and λ_2 are the roots of the equation $\lambda^2 - \mu_1 = 0$ and c_1, c_2 are two constants of integration. A similar interpretation applies for the other equations $\dot{\mathbf{x}} = P_i \mathbf{x}$.

2.4 Treating the Diagonal with Exponentials

We now have several methods to treat the off-diagonal part. Let us focus next on how to treat the case when A is diagonal and traceless. The system $\dot{\mathbf{x}} = \Lambda \mathbf{x}$, with Λ a diagonal matrix, has exact solution

$$x_{1,k} = \mathrm{e}^{h\lambda_{k,k}} x_{0,k}, \quad k = 1, \ldots, n. \tag{7}$$

In particular, when $A = \Lambda$ is traceless, (7) is volume-preserving.

2.5 Treating the Diagonal by a Rank-1 System

The next method is based on the following idea: set $\mathbf{d} = \operatorname{diag} A$, and denote $\mathbf{1} = [1, 1, \ldots, 1]^{\mathrm{T}}$. We write

$$A = \tilde{A} + (A - \tilde{A}), \quad \tilde{A} = \mathbf{1}\mathbf{d}^{\mathrm{T}}, \tag{8}$$

where $\tilde{A}$ is a rank-1 matrix, whose diagonal coincides with the diagonal of A, and $(A - \tilde{A})$ is a matrix with zero diagonal elements that we can treat with any of the methods for off-diagonal elements described earlier.

For $\dot{\mathbf{x}} = \tilde{A}\mathbf{x} = \mathbf{1}\mathbf{d}^{\mathrm{T}}\mathbf{x}$, we observe that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{d}^{\mathrm{T}}\mathbf{x} = \mathbf{d}^{\mathrm{T}}\dot{\mathbf{x}} = \big(\mathbf{d}^{\mathrm{T}}\mathbf{1}\big)\mathbf{d}^{\mathrm{T}}\mathbf{x} = 0$$

as $\mathbf{d}^{\mathrm{T}}\mathbf{1} = \operatorname{tr} A = 0$. Hence $\mathbf{d}^{\mathrm{T}}\mathbf{x}$ is constant and $\dot{\mathbf{x}} = \tilde{A}\mathbf{x}$ can be solved explicitly and exactly by a step of Forward Euler,

$$\mathbf{x}_1 = \mathbf{x}_0 + h\mathbf{1}\mathbf{d}^{\mathrm{T}}\mathbf{x}_0. \tag{9}$$

The matrix $\tilde{A}$ has the property that $\tilde{A}^2 = 0$ and thus $\exp(h\tilde{A}) = I + h\tilde{A}$.

The flow (9) is an example of a so-called shear. An ODE on $\mathbb{R}^n$ is a *shear* if there exists a basis of $\mathbb{R}^n$ (a linear change of coordinates) in which the ODE takes the form

$$\dot{y}_i = \begin{cases} 0, & i = 1, \ldots, k, \\ f_i(y_1, \ldots, y_k), & i = k+1, \ldots, n, \end{cases}$$

for some k. A diffeomorphism of $\mathbb{R}^n$ is called a shear if it is the flow of a shear. The advantage of splitting into shears is that their exact solution is computed by the Forward Euler method.

2.6 Splitting in $n + 1$ Shears

This method is a generalization of the method in Sect. 2.5. The splitting consists in decomposing the matrix A in (1) as a sum $n+1$ rank-1 matrices of the form $\mathbf{a}_i\mathbf{b}_i^{\mathrm{T}}$ with $\mathbf{a}_i, \mathbf{b}_i \in \mathbb{R}^n$. The volume-preservation condition then becomes $\mathbf{a}_i^{\mathrm{T}}\mathbf{b}_i = 0$, namely, that the vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ are orthogonal. The ODEs $\dot{\mathbf{x}} = \mathbf{a}_i\mathbf{b}_i^{\mathrm{T}}\mathbf{x}$ are shears.

Since the matrix A has $n^2 - 1$ free components, assuming that the $\mathbf{a}_i$ are given, for each $\mathbf{b}_i$ we have $n - 1$ free parameters (having taken into account orthogonality), which means that $n + 1$ shears are required:

$$A = \sum_{i=1}^{n+1} \mathbf{a}_i\mathbf{b}_i^{\mathrm{T}}, \quad \text{subject to} \quad \mathbf{a}_i^{\mathrm{T}}\mathbf{b}_i = 0. \tag{10}$$

The cases studied in [11] indicate that the vectors $\mathbf{a}_i$ which are on average best for all matrices should be chosen to be as widely or regularly spaced as possible. In this

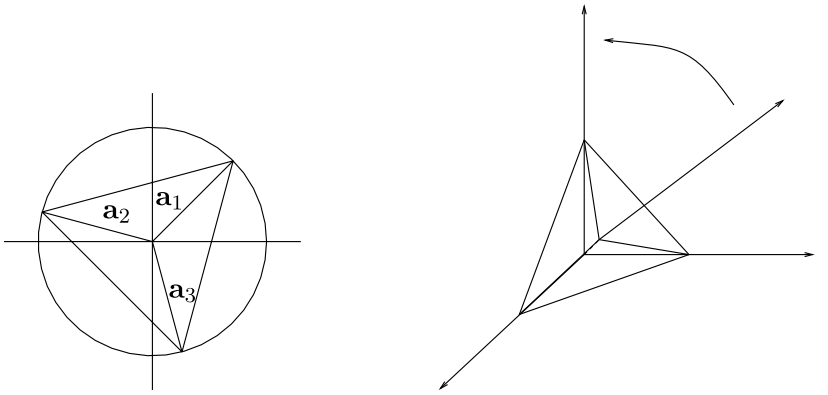

Fig. 1 The 2-simplex in $\mathbb{R}^2$ (*left*) and its construction (*right*)

case, a maximally symmetric configuration is possible, namely, the lines through the origin to the vertices of the n-simplex in $\mathbb{R}^n$ inscribed in S^{n-1}. These vectors $\mathbf{a}_i$ can be computed in the following manner. Start with the $n+1$ simplex in $\mathbb{R}^{n+1}$ with the vertices $\mathbf{e}_i$, the ith basis vector in $\mathbb{R}^{n+1}$. Thereafter, construct the vectors $\tilde{\mathbf{a}}_i$ in $\mathbb{R}^{n+1}$ as

$$\tilde{\mathbf{a}}_i = \mathbf{e}_i - \mathbf{b}_c, \quad i = 1, \ldots, n+1,$$

where $\mathbf{b}_c = [1/(n+1)]\mathbf{1}^{\mathrm{T}}$ is the baricentre of the simplex. Note that the $\tilde{\mathbf{a}}_i$'s are orthogonal to $\mathbf{b}_c$, hence they lie in an n-dimensional subspace of $\mathbb{R}^{n+1}$. To obtain the coordinates in $\mathbb{R}^n$, it is sufficient apply a rotation that maps $\mathbf{b}_c$ to the one of the axes, for instance $\mathbf{e}_1$. This can easily be achieved by a Householder reflection P. The vectors $P\tilde{\mathbf{a}}_i$ will now have zero in the first component, hence the remaining components give its coordinates in $\mathbb{R}^n$. Finally, these $n+1$ vectors are normalized to obtain the $\mathbf{a}_i$'s.

The procedure is illustrated in Fig. 1 for the 2-simplex.

To compute the $\mathbf{b}_i$'s, we multiply (10) by $\mathbf{a}_j^{\mathrm{T}}$ on the left and, denoting by C the matrix with entries $c_{i,j} = \mathbf{a}_i^{\mathrm{T}}\mathbf{a}_j = [(n+1)/n]\tilde{\mathbf{a}}_i^{\mathrm{T}}\tilde{\mathbf{a}}_j$, we obtain the system

$$C \begin{bmatrix} \mathbf{b}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{b}_{n+1}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{a}_{n+1}^{\mathrm{T}} \end{bmatrix} A. \tag{11}$$

The matrix C is singular, as the row sum $\sum_{j=1}^{n+1} c_{i,j} = [(n+1)/n]\tilde{\mathbf{a}}_i^{\mathrm{T}} \sum_{j=1}^{n+1} \tilde{\mathbf{a}}_j = 0$ because the sum of the $\tilde{\mathbf{a}}_j$ equals zero. Since the rows (and columns) of C coincide with the $\tilde{\mathbf{a}}_i$'s (up to a constant), the matrix C has rank n (as the $\mathbf{a}_i$ span $\mathbb{R}^n$). Thus, the $n \times n$ principal minor of C can be inverted to find the components of the vectors $\mathbf{b}_i$, which are defined up to an arbitrary component. The last component can thereafter be determined by imposing the orthogonality condition $\mathbf{a}_i^{\mathrm{T}}\mathbf{b}_i = 0$ for $i = 1, 2, \ldots, n+1$.

Given the splitting (10) of the matrix A, for $i = 1, \ldots n+1$, we solve the ODEs

$$\dot{\mathbf{x}} = \mathbf{a}_i \mathbf{b}_i^{\mathrm{T}} \mathbf{x}. \tag{12}$$

Each ODE is a shear. As in the previous section, $(\mathrm{d}/\mathrm{d}t)\mathbf{b}_i^{\mathrm{T}}\mathbf{x} = (\mathbf{b}_i^{\mathrm{T}}\mathbf{a}_i)\mathbf{b}_i^{\mathrm{T}}\mathbf{x} = 0$, hence each ODE is solved exactly by a step of Forward Euler.

3 Quadratic Volume-Preserving Vector Fields

3.1 Introduction

In this section we focus on quadratic volume-preserving vector fields,

$$\dot{x}_i = \sum_{\substack{j,k=1 \\ j\le k}}^{n} a_{i,j,k} x_j x_k, \quad i = 1, \ldots, n, \tag{13}$$

which, in short, we will denote by $\dot{\mathbf{x}} = \mathbf{A}(\mathbf{x}, \mathbf{x})$. The divergence-free condition becomes the set of n equations

$$\begin{aligned} 2a_{1,1,1} + a_{2,1,2} + \cdots + a_{n,1,n} &= 0, \\ 2a_{2,2,2} + a_{1,1,2} + \cdots + a_{n,2,n} &= 0, \\ &\vdots \\ 2a_{n,n,n} + a_{n,1,n} + \cdots + a_{n,n-1,n} &= 0, \end{aligned} \tag{14}$$

for the coefficients of the tensor $\mathbf{A}$. Also, in the quadratic case, we will talk about the diagonal part and the off-diagonal part of (13). By diagonal part, we intend all the terms that, at equation i, involve the variable x_i, corresponding to the set of coefficients $\{a_{i,j,k}, i, j, k = 1, \ldots, n,\ \ j \le k : j = i \vee k = i\}$. The off-diagonal part is the complement, $\{a_{i,j,k}, i = 1, \ldots, n,\ \ j \le k : j \ne i \wedge k \ne i\}$. From the definition of divergence, it is clear that only the coefficients of the diagonal part are involved in the divergence-free condition (14), therefore, a quadratic system with zero diagonal part is automatically divergence-free.

This section is organized as Sect. 2. We will first introduce methods for treating separately the diagonal part and the off-diagonal part (Sects. 3.2–3.5) and then the global methods (Sects. 3.6–3.7).

3.2 Splitting the Off-Diagonal Part Using Canonical Directions

As for the linear case, we start with the case when the quadratic system has zero diagonal part, that is, $a_{i,j,k} = 0$, whenever $j = i$ or $k = i$ for $i = 1, \ldots, n$.

The simplest method to deal with this system is a generalization of the method (4) for the linear case: we write the tensor $\mathbf{A}$ as the sum of tensors $\mathbf{R}_1, \ldots, \mathbf{R}_n$, where $r_{l;i,j,k} = 0$ for $i \ne l$, while $r_{l;i,j,k} = a_{i,j,k}$ for $i = l$, $l = 1, \ldots, n$. Then the differential equation $\dot{\mathbf{x}} = \mathbf{R}_l(\mathbf{x}, \mathbf{x})$ reduces to

$$\begin{aligned} \dot{x}_m &= 0, \quad m \ne l, \\ \dot{x}_l &= \sum_{\substack{j,k=1 \\ j\le k}}^{n} a_{l,j,k} x_j x_k \end{aligned} \tag{15}$$

and can be solved exactly by a step of Forward Euler. See also [8] for a generalization to arbitrary volume-preserving vector fields for which $\partial f_i/\partial x_i = 0, i = 1, \ldots, n$.

3.3 Treating the Off-Diagonal Part by Lower-Triangular Systems

Assume $a_{i,j,k} = 0$, whenever $j = i$ or $k = i$ for $i = 1, \ldots, n$. We start with the following result.

Proposition 2 *Consider the divergence-free differential equation*

$$\begin{aligned} \dot{x}_{i_1} &= 0, \\ \dot{x}_{i_2} &= f_{i_2}(x_{i_1}), \\ &\vdots \\ \dot{x}_{i_n} &= f_{i_n}(x_{i_1}, \ldots, x_{i_{n-1}}), \end{aligned} \tag{16}$$

where $i_1, i_2, \ldots, i_n$ is any permutation of the indices $\{1, 2, \ldots, n\}$. Any Runge–Kutta method applied to (16) is volume-preserving.

The proof of the above result is very similar to that of Proposition 1. The Jacobian of the method is of the form I, the identity matrix, plus a strictly lower triangular matrix and has hence determinant equal to 1. In particular, also in this case, we have that the Backward Euler can be implemented explicitly and it can be combined with Forward Euler to obtain higher-order volume-preserving schemes.

We will call systems of type (16) *strictly triangular systems*, in analogy with the linear case. However, it is easily observed, by counting the number of free parameters, that it is not generally possible to split any quadratic off-diagonal part into two strictly triangular systems only.

Let us introduce the following simplifying notation. By writing the column

$$\begin{matrix} i_1 \\ i_2 \\ \vdots \\ i_n \end{matrix} \tag{17}$$

we intend a differential equation of type (16). How to choose the coefficients $a_{i,j,k}$ to be put in the f_{l_m}? If, for instance, $i_1 = 1,\ i_2 = 2, \ldots, i_n = n$, it is clear that in the second equations we can put only $a_{2,1,1}x_1^2$, as it is the only off-diagonal term including only x_1. Similarly, in the third equation we can put $a_{3,1,1}x_1^2$, $a_{3,1,2}x_1x_2$, $a_{3,2,2}x_2^2$, and so on. In the last equation, we can put all the $a_{n,j,k}$-terms as $j, k < n$. Hence, the terms defining $\dot{x}_n$ are taken care of.

In general, the splitting can be described by s columns like (17), each column corresponding to a permutation of the indices $\{1, \ldots, n\}$.

What is the minimal number of terms s in which we need to split the vector field (13) so that each elementary vector field is of the form (16) and can be solved explicitly by FE/BE? Clearly, $s \leq n$, as the n systems (15) are strictly triangular (up to a permutation of the indices).

The minimization problem turns out to be quite difficult. Below we discuss an algorithm that does not produce the optimal solution but, nevertheless, gives a number s of splittings growing like $\sqrt[3]{n}$ instead of n for n large. We start with an example and from that we will describe a general algorithm. In the columns below we show how to split an eight-dimensional quadratic vector field (13) with zero diagonal part, in $s = 4$ terms,

$$\left|\begin{array}{c|c|c|c} \vdots & \vdots & \vdots & \vdots \\ 3 & 4 & 1 & 2 \\ 2 & 1 & 4 & 3 \\ 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{array}\right|$$

where the dots mean that each column is filled with the remaining indices, in arbitrary order. Note that 5, 6, 7, 8 appear at the bottom, which means that we can include all their right-hand side terms in (13). Now focus on the index 1. In the first column all the indices except 5 occur above 1, here we can put all the right-hand side terms of type $a_{1,i,j}x_ix_j$ for $\dot{x}_1$, with $i, j \neq 5$. In the second column, 5 comes somewhere above 1, while 2, 6 are below, thus in the second vector field, for the x_1 variable, we can put all the missing terms of type $a_{1,i,j}$, where i or j equal 5, except those of type $a_{1,2,5}x_2x_5, a_{1,5,6}x_5x_6$. In the third column, 5, 6, 2 all come above 1, hence the remaining terms $a_{1,2,5}x_2x_5, a_{1,5,6}x_5x_6$ can be placed here. A similar procedure yields the other indices. Given the term $a_{i,j,k}$, we say that a column is *admissible* for $a_{i,j,k}$ if, in that column, j, k are above i. For each $a_{i,j,k}$ there might be several admissible columns. In this example, the term $\dot{x}_1 = a_{1,8,8}x_8^2$ can go in the first, second or third column as the indices $j = 8, k = 8$ are above the index 1 in each of these columns. Thus, the first, second and third columns are admissible for $a_{1,8,8}$. To make the choice unique, we put the term $a_{i,j,k}$ in the first admissible column. Thus, in our example, the term $\dot{x}_1 = a_{1,8,8}x_8^2$ goes in the first column. Another possibility could be to average the term $a_{i,j,k}$ among all its admissible columns. We will not consider this choice as it would increase the computational cost and the complexity of the method.

The problem can be formalized as follows.

Problem 1 *Find an integer table $P_{n,s}$ with s columns, each containing a permutation of $\{1, \ldots, n\}$, such that for each triplet of distinct indices, i, j, k, it is true that $i \prec j, k$ in at least one column, where the symbol $\prec$ means 'is below'.*

Although the optimal solution of this problem is not known, the algorithm below is a relatively simple solution with $n = \binom{s}{3} + s$, and thus $s = \mathcal{O}(n^{1/3})$. Our construction is based on the following result.

Lemma 3 *Given an $n \times s$ integer table P, where each column is a permutation of $\{1, \ldots, n\}$, if for any given integer i there exist three columns such that for any $j \neq i$ it is true that $j \prec i$ in at most one of these three columns, then P solves Problem 1.*

Proof Given three distinct indices i, j, k, then $j \prec i$ or $k \prec i$ in at most two of the three columns. Thus in the third column $i \prec j, k$. □

Algorithm 1 We construct the permutation table $P_{n,s}$ as follows:

1. Create a partial table P_s of size $\binom{s-1}{2} \times s$, containing exactly three copies of each of the integers $\{1, 2, \ldots, \binom{s}{3}\}$ by the following induction:
   (a) $P_3 = \boxed{1\ 1\ 1}$.
   (b) For $s > 3$ create P_s from P_{s-1} as follows:
      (i) For $1 \le i < s$, fill column i of P_s with column i of P_{s-1} lifted up $i-1$ positions. In Matlab/Octave notation:

$$k = \text{size}(P_{s-1}, 1); \quad \text{for } i = 1 : s-1,\ P_s(i : i+k-1, i) = P_{s-1}(:, i);\ \text{end}$$

      (ii) P_s is now defined, except in three regions: A triangular upper left part, a triangular lower right part and the rightmost column. Each of these regions are filled with the ordered list of integers $\left\{\binom{s-1}{3} + 1, \ldots, \binom{s}{3}\right\}$ as follows:
      - The upper left triangle is filled columnwise from left to right and from bottom to top within each column.
      - The lower right triangle is filled rowwise from bottom to top and from left to right within each row.
      - The rightmost column of P_s is filled from the top down.

   The first three P_3, P_4 and P_5 are given as

$$\boxed{1\ 1\ 1} \quad \rightarrow \quad \begin{array}{|c|c|c|c|}\hline 3 & 4 & 1 & 2 \\ \hline 2 & 1 & 4 & 3 \\ \hline 1 & 2 & 3 & 4 \\ \hline \end{array} \quad \rightarrow \quad \begin{array}{|c|c|c|c|c|}\hline 7 & 9 & 10 & 2 & 5 \\ \hline 6 & 8 & 1 & 3 & 6 \\ \hline 5 & 4 & 4 & 4 & 7 \\ \hline 3 & 1 & 3 & 10 & 8 \\ \hline 2 & 2 & 8 & 9 & 9 \\ \hline 1 & 5 & 6 & 7 & 10 \\ \hline \end{array}$$

2. Create $P_{n,s}$ from P_s by adding an additional row on the bottom containing the integers $\{\binom{s}{3} + 1, \ldots, \binom{s}{3} + s\}$, and add the remaining integers on top of each column in an arbitrary order so that the columns of $P_{n,s}$ are permutations.

Lemma 4 *$P_{n,s}$ produced by Algorithm 1 solves Problem 1.*

Proof Consider the partial table P_s produced in Step 1. We see that each integer i appears in three different columns. By construction we have arranged the integers so that each of the possible $\binom{s}{3}$ ways that three columns can be selected from s are found exactly once among the integers $1, \ldots, \binom{s}{3}$. Thus any pair of indices i, j are found together in at most two columns. By induction we show that for any $j \neq i$, the relation $j \prec i$ holds in at most *one* column of P_s. This is clearly so for $s = 3$. For an arbitrary s we assume that it is true for P_{s-1}. Now pick two different indices $i, j \in P_s$.

- If $i, j \in P_{s-1}$, then the induction hypothesis yields that $j \prec i$ in at most one column.

- If $i \in P_{s-1}$ and $j \in P_s \setminus P_{s-1}$, then $j \prec i$ in at most one column, since each of the new integers appears only once below P_{s-1} in P_s. Similarly, if $j \in P_{s-1}$ and $i \in P_s \setminus P_{s-1}$ then $j \prec i$ in at most one column, since each of the new integers appears only once above P_{s-1} in P_s.
- Finally, if $i, j \in P_s \setminus P_{s-1}$, then i and j appear together in the final column s and in at most one of the columns $1, \ldots, s-1$. However, in the first $s-1$ columns the indices are sorted upwards, while they are sorted downwards in the last. Thus $j \prec i$ holds in at most one column.

Now, from Lemma 3 it follows that $P_{n,s}$ solves Problem 1; the addition of s unique integers at the bottom does not destroy the three-column property for the integers $1, \ldots, \binom{s}{3}$. The final s integers on the bottom obviously satisfy Problem 1. □

It should be noted that our construction is the tightest possible based on the three-column property, since we have exhausted all possible selections of three columns. If two integers i, j share the same selection of three columns, we must have either $i \prec j$ or $j \prec i$ in at least two of these. However, there are shorter solutions to Problem 1 which do not satisfy the three-column property.

3.4 Treating the Diagonal Part by Exponentials

Let us consider next the case when the diagonal part of $\mathbf{A}$ is nonzero, while the off-diagonal part is zero. Unlike the linear case, the diagonal part of the system is not necessarily integrable. It must also be split. A possible splitting of the diagonal part has to take into account the divergence-free conditions (14), as breaking the equations would result in a method that is not volume-preserving.

Our idea is to split the diagonal part of the vector field in such a way that each split term obeys one of the n conditions in (14). As an example, for the first condition, we obtain the divergence-free vector field

$$\begin{aligned} \dot{x}_1 &= a_{1,1,1} x_1^2, \\ \dot{x}_2 &= a_{2,1,2} x_1 x_2, \\ &\vdots \\ \dot{x}_n &= a_{n,1,n} x_1 x_n. \end{aligned} \tag{18}$$

The first differential equation involves the variable x_1 only and can be solved exactly in the interval $[t_0, t]$,

$$x_1(t) = \frac{x_1(t_0)}{1 - a_{1,1,1} x_1(t_0)(t - t_0)}.$$

Thereafter, each of the other equations is with variable coefficients but linear, and can be also solved exactly,

$$x_i(t) = x_i(t_0) \mathrm{e}^{a_{i,1,i} \int_{t_0}^{t} x_1(\tau) \mathrm{d}\tau} = x_i(t_0) \mathrm{e}^{a_{i,1,i} f_1(x_1(t_0), t)},$$

where

$$f_1(x_1(t_0),t)=\begin{cases}-\frac{1}{a_{1,1,1}}\ln(1-a_{1,1,1}x_1(0)(t-t_0)) & \text{if } a_{1,1,1}\neq 0,\\ (t-t_0)x_1(0) & \text{if } a_{1,1,1}=0.\end{cases} \tag{19}$$

Similar procedure for the terms involved in the other conditions, for a total of n vector fields to treat the diagonal part.

3.5 Treating the Diagonal Part by Two Shears

By a direct count of free parameters, McLachlan and Quispel [11] conjecture that a d-degree divergence-free polynomial vector field can be split in $n+d$ terms, each of them a function of $n-1$ variables ($n-1$ planes in $\mathbb{R}^n$). In coordinates, one chooses an orthonormal basis $\mathbf{a}_1,\ldots,\mathbf{a}_{n-1}$ for the $n-1$ plane and completes it to an orthonormal basis of $\mathbb{R}^n$ (or vice versa, can choose a direction vector $\mathbf{a}_n$ in $\mathbb{R}^n$ and find a basis for a plane perpendicular to $\mathbf{a}_n$). When $d=2$ (quadratic case), each split vector field (among the $n+2$ ones) can be written in the form

$$\dot{\mathbf{x}}=\mathbf{a}_n g\big(\mathbf{a}_1^{\mathrm{T}}\mathbf{x},\ldots,\mathbf{a}_{n-1}^{\mathrm{T}}\mathbf{x}\big),\quad \mathbf{a}_i^{\mathrm{T}}\mathbf{a}_j=\delta_{i,j},\quad i,j=1,\ldots,n,$$

where $\delta_{i,j}$ is the Kronecker delta, and g is a scalar function. By choosing $\mathbf{a}_n$ as any of the elementary unit vectors, we recover the method in Sect. 3.2 to treat the off-diagonal part. It remains to find two vectors $\mathbf{a},\mathbf{b}$ and their complements to an orthogonal basis, so that

$$\dot{\mathbf{x}}=\mathbf{a}g_1\big(\mathbf{a}_1^{\mathrm{T}}\mathbf{x},\ldots,\mathbf{a}_{n-1}^{\mathrm{T}}\mathbf{x}\big)+\mathbf{b}g_2\big(\mathbf{b}_1^{\mathrm{T}}\mathbf{x},\ldots,\mathbf{b}_{n-1}^{\mathrm{T}}\mathbf{x}\big) \tag{20}$$

can take care of the diagonal elements. By a simple count of free parameters, this should be possible: $\mathbf{a},\mathbf{a}_1,\ldots,\mathbf{a}_{n-1}$ and $\mathbf{b},\mathbf{b}_1,\ldots,\mathbf{b}_{n-1}$ define two orthogonal basis of $\mathbb{R}^n$ (orthogonal matrices), and each of the matrices can be determined in terms of $n(n-1)/2$ parameters, for a total of $n(n-1)=n^2-n$ free parameters, which corresponds to the number of free parameters for the diagonal part as one has n^2 coefficients and n constraints arising from volume preservation.

Lemma 5 ([11]) *Consider the differential equation*

$$\dot{\mathbf{x}}=g\big(\mathbf{a}_1^{\mathrm{T}}\mathbf{x},\ldots,\mathbf{a}_m^{\mathrm{T}}\mathbf{x}\big)\mathbf{a},\quad \mathbf{x}(0)=\mathbf{x}_0, \tag{21}$$

where $\mathbf{a}_1,\ldots,\mathbf{a}_m,\mathbf{a}\in\mathbb{R}^n$ *and* $g:\mathbb{R}^m\to\mathbb{R}$ *is a scalar function. If*

$$\mathbf{a}_i^{\mathrm{T}}\mathbf{a}=0,\quad i=1,2,\ldots,m,$$

then:

(i) (21) *is divergence-free*; *and*
(ii) *the functions* $\mathbf{a}_i^{\mathrm{T}}\mathbf{x}$ *are independent of time*;

hence (21) *has exact solution*

$$\mathbf{x}(t)=\mathbf{x}_0+tg\big(\mathbf{a}_1^{\mathrm{T}}\mathbf{x}_0,\ldots,\mathbf{a}_m^{\mathrm{T}}\mathbf{x}_0\big)\mathbf{a}. \tag{22}$$

Proof In the hypotheses of the lemma, we have

$$\operatorname{div} f = \sum_{i=1}^{m} (\mathbf{a}^{\mathrm{T}}\mathbf{a}_i) g_i = 0, \quad g_i = \frac{\partial}{\partial y_i} g(y_1, \ldots y_m),$$

hence (i) follows. Condition (ii) follows as easily by observing that $(\mathrm{d}/\mathrm{d}t)\mathbf{a}_i^{\mathrm{T}}\mathbf{x} = (\mathbf{a}_i^{\mathrm{T}}\mathbf{a}) g(\mathbf{a}_1^{\mathrm{T}}\mathbf{x}, \ldots, \mathbf{a}_m^{\mathrm{T}}\mathbf{x}) = 0$. □

Note that the system (20) is a generalization of (8).

We remark that the vectors $\mathbf{a}_i$ in the above lemma need not be linearly independent. The important argument is that they are orthogonal to the direction of advancement $\mathbf{a}$ and that they should span an $(n-1)$-dimensional space. Therefore, we use the ansatz

$$\dot{\mathbf{x}} = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{bmatrix} \sum_{i<j} \alpha_{i,j} (A_j x_i - A_i x_j)^2, \tag{23}$$

where $A_j x_i - A_i x_j = (A_j \mathbf{e}_i - A_i \mathbf{e}_j)^{\mathrm{T}}\mathbf{x}$, the vectors $A_j \mathbf{e}_i - A_i \mathbf{e}_j$ all being orthogonal to $[A_1, \ldots, A_n]^{\mathrm{T}} = \sum A_k \mathbf{e}_k$.

The $\alpha_{i,j}$ in (23) constitute $n(n-1)/2$ coefficients, that are going to determine the first vector field in (20). We also introduce the coefficients $\beta_{i,j}$ corresponding to the second vector field. Matching coefficients, we obtain $n(n-1)/2$ sets of equations of the type

$$\begin{bmatrix} A_i^2 A_j & B_i^2 B_j \\ A_i A_j^2 & B_i B_j^2 \end{bmatrix} \begin{bmatrix} \alpha_{i,j} \\ \beta_{i,j} \end{bmatrix} = -\frac{1}{2} \begin{bmatrix} a_{i,i,j} \\ a_{j,i,j} \end{bmatrix}, \quad 1 \le i < j \le n, \tag{24}$$

which determine the values of $\alpha_{i,j}, \beta_{i,j}$, as long as the determinant $A_i A_j B_i B_j (A_i B_j - A_j B_i)$ of the coefficient matrices is nonzero. In particular, it follows that the A_k and B_k cannot generally be zero.

As an example, consider the volume-preserving system

$$\dot{x}_1 = \frac{a}{2} x_1^2 - b x_1 x_2,$$

$$\dot{x}_2 = -a x_1 x_2 + \frac{b}{2} x_2^2,$$

and assume that $[A_1, A_2]^{\mathrm{T}} = [1, 1]^{\mathrm{T}}$, and $[B_1, B_2]^{\mathrm{T}} = [1, -1]$. If $a = 3, b = 2$, the desired coefficients $\alpha_{1,2}, \beta_{1,2}$ then become $\alpha_{1,2} = 5/4$ and $\beta_{1,2} = 1/4$, and the system can be written in the form

$$\dot{\mathbf{x}} = \frac{5}{4} \begin{bmatrix} 1 \\ 1 \end{bmatrix} (x_1 - x_2)^2 + \frac{1}{4} \begin{bmatrix} 1 \\ -1 \end{bmatrix} (-x_1 - x_2)^2 = \begin{bmatrix} \frac{3}{2} x_1^2 - 2 x_1 x_2 + \frac{3}{2} x_2^2 \\ -3 x_1 x_2 + x_2^2 + x_1^2 \end{bmatrix},$$

so that each of the two split terms can be exactly integrated as in (22). Note that this procedure introduces off-diagonal terms ($3/2 x_2^2$ for $\dot{x}_1$ and x_1^2 for $\dot{x}_2$) one has to account for when treating the off-diagonal part. This is in fact analogous to what happens for the rank-1 splitting described in Sect. 2.5.

3.6 Splitting in Lower Triangular Systems by Orthogonal Changes of Variables

As we already know, a system of the form

$$\dot{y}_i = \sum_{j \le k} c_{i,j,k} y_j y_k, \tag{25}$$

where $c_{i,j,k}$ defines a strictly lower triangular tensor (i.e. $c_{i,j,k} = 0$ for $i \le j, k$), and can be easily solved in a volume-preserving manner by either Forward Euler or Backward Euler. Setting $\mathbf{y} = Q\mathbf{x}$, where Q is an orthogonal matrix, we obtain

$$\dot{x}_i = \sum_{l,m} \sum_{p,j,k} Q_{i,p} c_{p,j,k} Q_{j,l} Q_{k,m} x_l x_m,$$

thus,

$$\sum_{p,j,k} Q_{i,p} c_{p,j,k} Q_{j,l} Q_{k,m} = a_{i,l,m},$$

a linear system in the unknowns $c_{p,j,k}$. For a strictly triangular system, there are $n_c = \frac{1}{6} n(n-1)(n+1)$ parameters $c_{p,j,k}$, while the total number of free parameters $a_{i,j,k}$ is $n_a = N(n,d)n - N(n,d-1) = \frac{1}{2} n(n^2+n-2)$, where $N(n,d) = \binom{n+d-1}{d}$ and $d = 2$ [11]. By taking the ratio $n_a/n_c = 3(n+2)/(n+1)$, we see that at most $s = 4$ strictly triangular systems (25) are needed, yielding the linear system

$$\sum_{s=1}^{4} \sum_{p,j,k} Q^{[s]}_{i,p} c^{[s]}_{p,j,k} Q^{[s]}_{j,l} Q^{[s]}_{k,m} = a_{i,l,m}$$

for the unknowns $c^{[s]}_{p,j,k}$. Several trial numerical experiments for dimension n up to 12 indicate that the system has a solution for a given set of orthogonal matrices $Q^{[s]}$, $s = 1, \ldots, 4$. As a simplifying condition, one of the matrices can be chosen to be the identity matrix. To get stable numerical methods, the matrices $Q^{[s]}$ should be chosen so to minimize the conditioning of the basis.

For the practical implementation of the method, we solve for four systems $\dot{\mathbf{x}} = \mathbf{A}^{[s]}(\mathbf{x}, \mathbf{x})$, where the (i,l,m) component of $\mathbf{A}^{[s]}$ is $a^{[s]}_{i,l,m} = \sum_{p,j,k} Q^{[s]}_{i,p} c^{[s]}_{p,j,k} Q^{[s]}_{j,l} \times Q^{[s]}_{k,m}$, by Forward Euler or Backward Euler.

3.7 Optimal $n+2$ Shears

We know from Sect. 3.5 that it is possible to split any given quadratic volume-preserving vector fields in the sum of $n+2$ shears. To do so, we can either choose the n canonical directions and two extra directions to account for the diagonal part (corresponding to the methods described in Sects. 3.2 and 3.5, or, another possibility is to take some other $n+2$ shears. These are of the form (23) with the coefficients α_{ij} replaced by $\alpha_{ij}/\sqrt{2}$ for rotational invariance and are solved exactly by a single step of Forward Euler.

As discussed in [11], these directions should be chosen so to minimize the conditioning of the basis.

Good Grassmannian packings [2] are expected to give bases with good condition numbers. For linear vector fields, we have used the $n+1$ vertices of a regular n-simplex. Already, for $d=2$, and arbitrary n, optimal packings are not always analytically known. For $n=3$, the regular polyhedra only give maximally symmetric sets of 3, 4, 6 or 10 lines through the origin—not 5 (although one could use the six diameters of the icosahedron, giving a nonunique splitting). For higher dimensions, vectors with maximal angle separation are computed numerically by means of nonlinear optimization techniques [2].

It is not clear that maximal angle separation (Grassmannian packing) gives the optimal $n+2$ directions for the shears in the quadratic problem, however, these are good starting points for a nonlinear least squares search, where the optimal solution is the one that minimizes the conditioning of the basis.

4 Numerical Examples

4.1 Comparison of Second-Order Methods for the Linear Case

To test the performance of the methods, we compare them on 2000 divergence-free linear systems for a problem of dimension $n=10$. The coefficients are chosen at random (Gaussian distribution) in the interval $[-1,1]$. Then we set the trace to zero by subtracting the quantity $1/n\,\mathrm{tr}(A)I$, where I is the $n\times n$ identity matrix. Finally, matrix A is normalized with respect to the 2-norm. Similarly, the initial condition is chosen at random on the sphere of radius one. The integration is performed over a short interval $[0,2]$ with different stepsizes $h_l=1/2^l$, where $l=0,1,2,\ldots,7$, and the 2-norm of the error is evaluated (for each stepsize and for each experiment). Finally, for each l, the error is averaged by taking the arithmetic mean over the samples. We compare several methods described in Sect. 2. The basic first-order methods are composed with their adjoints to obtain second-order schemes. Typically, the contribution of the diagonal part is put in the middle of the composition, as the methods in Sects. 2.4 and 2.5 solve it exactly and two half steps can be subsumed in a single step. This is not the case for the off-diagonal parts, where subsuming in a single step a half step with an FE (resp., BE) and half step with its adjoint BE (resp., FE) would yield an implicit method.

In more detail, we compare the method DEXP + LTS (strictly triangular systems, see Sect. 2.1, plus diagonal treated with exponentials, see Sect. 2.4), which costs $4n^2$ addition/mult. for the off-diagonal part and $n(E+2)$ for the diagonal part, where E denotes the cost of an exponential; the method DS + LTS (strictly triangular systems, Sect. 2.1, and rank-1 approximation of the diagonal, Sect. 2.5), whose cost differs from above as the diagonal part costs $3n$ additions/mult.; the method DEXP + N shears (the n canonical shears, Sect. 2.2, plus diagonal treated with exponentials, Sect. 2.4), costing $4n^2$ add./mult. for the off-diagonal and $n(E+2)$ for the diagonal; the splitting (N + D) shears in $n+1$ shears described in Sect. 2.6, which costs all together $8n^2+4n$ add./mult. and, finally, the method SYMPOL (symmetric polar splitting, Sects. 2.3 and 2.4), which costs $5\frac{1}{2}n^2$ add./mult. as the coefficients μ_j

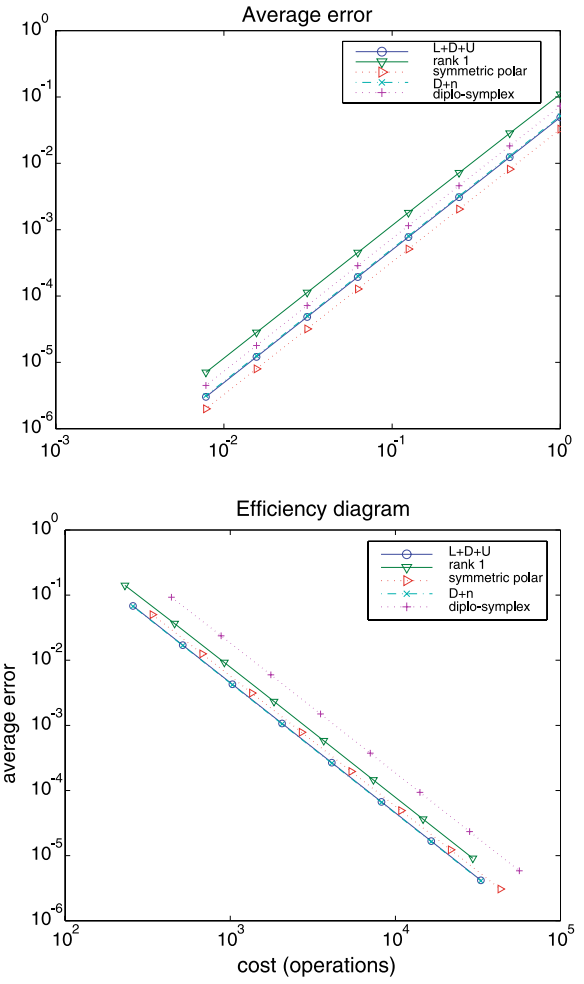

Fig. 2 *Top*: Average error over 2000 random linear problems. x-axis: step size, logarithmic scale; y-axis: average global error, logarithmic scale. *Bottom*: Efficiency diagram. x-axis: cost in number of operations, logarithmic scale; y-axis: average global error, logarithmic scale

need be computed only once, plus $2n(T+1)$, where T is the cost of a trigonometric function (see also [7]).

Since even a (non-volume-preserving) Forward Euler step applied to a linear system requires $2n^2$ operations, we see that the cheapest second-order volume-preserving methods cost about two Forward Euler steps; there is essentially no cost associated with requiring volume-preservation.

The factors E, T have been computed numerically on a Mac Power Processor (by taking a random matrix of dimension 1000×1000, applying the respective transformation elementwise and then comparing the elapsed time with that of sums of two matrices of the same dimension) and correspond to $E \approx 3.74$ and $T \approx 2.13$. Clearly, these factors are machine-dependent.

4.2 A Nine-Dimensional Lorenz System

To illustrate the behaviour of the methods for quadratic divergence-free methods, we apply them to approximate numerically a nine-dimensional Lorenz system [15],

which arises from the study of three-dimensional cells with square planform in dissipative Rayleigh–Bénard convection.

The system reads

$$\begin{aligned}
\dot{x}_1 &= -\sigma b_1 x_1 - \sigma b_2 x_7 - x_2 x_4 + b_3 x_3 x_5 + b_4 x_4^2,\\
\dot{x}_2 &= -\sigma x_2 - \sigma x_9/2 + x_1 x_4 - x_2 x_5 + x_4 x_5,\\
\dot{x}_3 &= -\sigma b_1 x_3 + \sigma b_2 x_8 - b_3 x_1 x_5 + x_2 x_4 - b_4 x_4^2,\\
\dot{x}_4 &= -\sigma x_4 + \sigma x_9/2 - x_2 x_3 - x_2 x_5 + x_4 x_5,\\
\dot{x}_5 &= -\sigma b_5 x_5 + x_2^2/2 - x_4^2/2,\\
\dot{x}_6 &= -b_6 x_6 + x_2 x_9 - x_4 x_9,\\
\dot{x}_7 &= -r x_1 - b_1 x_7 + 2 x_5 x_8 - x_4 x_9,\\
\dot{x}_8 &= r x_3 - b_1 x_8 - 2 x_5 x_7 + x_2 x_9,\\
\dot{x}_9 &= -r x_2 + r x_4 - x_9 - 2 x_2 x_6 - x_2 x_8 + 2 x_4 x_6 + x_4 x_7,
\end{aligned} \tag{26}$$

where $r = R/R_c$ is the *reduced Rayleigh number*, R being the Rayleigh number and $R_c = 27/4$ the critical Rayleigh number. The constant parameters b_i, measuring the geometry of the square cell, are given as

$$b_1 = \frac{4(1+a^2)}{1+2a^2}, \qquad b_2 = \frac{1+2a^2}{2(1+a^2)}, \qquad b_3 = \frac{2(1-a^2)}{1+a^2},$$
$$b_4 = \frac{a^2}{1+a^2}, \qquad b_5 = \frac{8a^2}{1+2a^2}, \qquad b_6 = \frac{4}{1+2a^2},$$

where $a = 1/2$ is the wave number in the horizontal direction. The parameter $\sigma = 1/2$ and $r = 14.22$ are the same as in [15].

Introducing the vector $\mathbf{x} = [x_1, \ldots, x_9]^{\mathrm{T}}$, the system can be written in the form

$$\dot{\mathbf{x}} = L\mathbf{x} + \mathbf{Q}(\mathbf{x}, \mathbf{x}), \tag{27}$$

namely, a linear term and a quadratic term. The quadratic term is divergence-free, while the linear term has negative divergence, i.e. the volume of the phase space is contracted at a constant rate.

A second-order numerical integrator for (27) contracting volume at the correct rate is

$$\Phi_h = \Phi_{h/2}^{[L]} \circ \Phi_h^{[\mathbf{Q}]} \circ \Phi_{h/2}^{[L]},$$

where $\Phi_\tau^{[L]}$ is a self-adjoint integrator for the linear part contracting volume at the correct rate, while $\Phi_\tau^{[\mathbf{Q}]}$ is a second-order volume-preserving method for the quadratic part.

As for as the linear part is concerned, it can easily be observed that it can be solved in practice exactly, as the variables have a weak coupling: x_5 and x_6 are independent from the other variables, x_1 is coupled with x_7, x_3 with x_8 and, finally, x_9 is coupled with x_2 and x_4 and vice versa. Therefore we will focus our attention on the quadratic part only.

The methods for the quadratic case are compared in Fig. 3. The basic methods are composed with their adjoints to obtain second-order schemes.

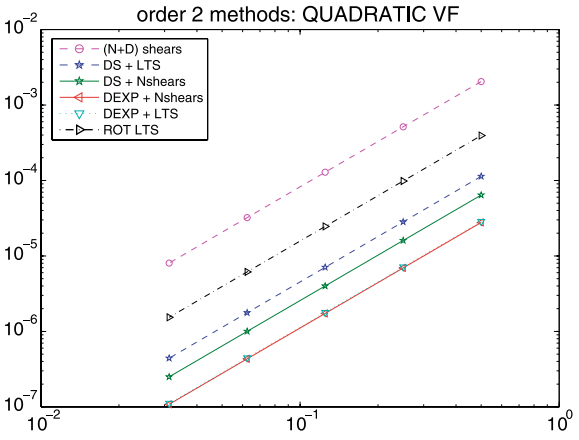

Fig. 3 Errors versus stepsize at $T = 2$ for the quadratic part of the Lorenz 9D system

For each term of the quadratic vector field, we have about three operations (two multiplications and one addition). There are about $\frac{1}{2}n^3$ off-diagonal terms and n^2 diagonal terms. Hence, the cost of the methods will be dominated by the treatment of the off-diagonal terms.

In more detail, we have: the methods DEXP + LTS and DS + LTS (strictly triangular systems plus diagonal treated with exponentials and diagonal shear, respectively), that cost $3n^3$ operations for the off-diagonal part and $(3 + E)n^2$ and $4n^2$ for the diagonal part respectively; the methods DEXP + N shears and DS + N shears (n canonical shears plus diagonal treated with exponentials and diagonal shears, respectively), that have exactly the same cost as the methods above; the method (N + D) shears ($n + 2$ shears), costing $8n^3$ operations and, finally, the method ROTLTS of rotations into strictly triangular shears, amounting to $4n^3$ operations (provided that the rotations are combined with the coefficients before the implementation).

All the methods attain second order (as expected), and are more or less comparable as far as the error is concerned. We have made no effort in optimizing the methods to the particular problem in question, for instance, in order to reduce the number of splitting terms due to the sparsity of the coefficients.

5 Conclusions

Numerical examples for both the linear case and the quadratic case indicate that methods that treat the diagonal with exponentials have an edge on the methods that split the diagonal part in shears. This can be partly explained by observing that the former methods are exact on each basis element for the basis of linear and quadratic divergence-free vector fields, whereas the methods that treat the diagonal with shears are not. The effect of a worse accuracy is evident in Fig. 4 for long time integration. The methods of splitting $n + d$ shears and rotations in lower triangular matrices display a much larger error in the individual trajectories, while the remaining methods show a solution much closer to the original one. Methods based on $n + d$ shears are on average twice more expensive than the other basic methods, although they

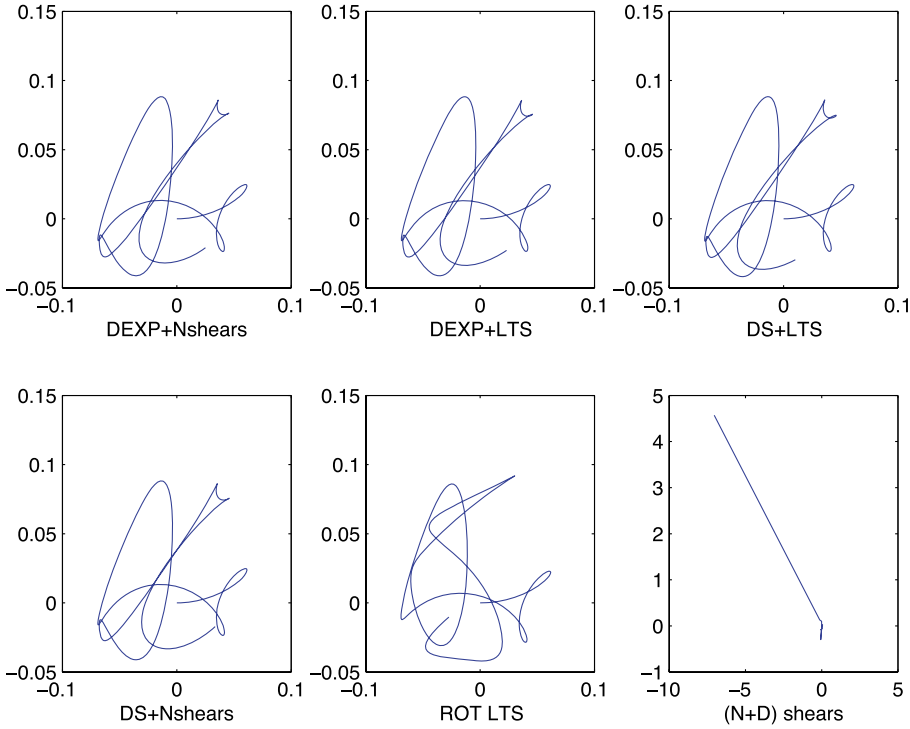

Fig. 4 Plots of the sixth component versus the seventh component of the quadratic part of the Lorenz 9D system in the interval $[0, 150]$ with integration step $h = 1/2$ and initial condition $x_i(0) = 1, i = 1, 2, 3, 9$ and $x_i(0) = 0$ otherwise. The plots are arranged from best (*top left*) to worst (*bottom right*). The method (N + D) shears diverges around $t = 90$

have the advantage that they can be implemented using only arithmetic operations, while some other methods require the computation of transcendental and trigonometric functions—though also these methods use only arithmetic operations on the computational-heavy part of the system (off-diagonal part). It is also important to say that both the methods based on rotations and on $n + d$ shears involve an extensive and costly nonlinear optimization step to find a good conditioned basis, and are therefore less attractive for short time integrations and problems of large dimension.

Acknowledgements We are very grateful to Drs Yajuan Sun and Will Wright for many useful contributions.

This research was supported by the Australian Research Council and by the Likestillingsmidlene from the University of Bergen. We would like to thank the Departments of Mathematics of the University of Bergen and La Trobe University for their hospitality.

References

1. C. J. Budd and A. Iserles, Geometric integration: Numerical solution of differential equations on manifolds, *Philos. Trans. Roy. Soc. A* **357** (1999), 945–956.

2. J. H. Conway, R. H. Hardin, and N. J. Sloan, Packing lines, planes, etc.: packing in Grassmannian spaces, *Experiment. Math.* **5**(2) (1996), 139–159.
3. P. Chartier and A. Murua, Preserving first integrals and volume forms of additively split systems, *IMA J. Numer. Anal.* **27**(2) (2007), 381–405.
4. A. J. Dragt and D. T. Abell, Symplectic maps and computations of orbits in particle accelerators, in *Fields Institute Communications*, Vol. 10, American Mathematical Society, Providence, 1996, pp. 59–85.
5. E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration*, Springer Series in Computational Mathematics, Vol. 31, Springer, Berlin, 2002.
6. A. Iserles, G. R. W. Quispel, and P. S. P. Tse, B-series methods cannot be volume-preserving, *BIT* **47**(2) (2007), 351–378.
7. A. Iserles and A. Zanna, *Efficient Computation of the Matrix Exponential by Generalized Polar Decompositions*, Technical Report NA2002/09, University of Cambridge, 2002.
8. F. Kang and Z.-J. Shang, Volume-preserving algorithms for source-free dynamical systems, *Numer. Math.* **71** (1995), 451–463.
9. B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge Monographs on Applied and Computational Mathematics, Vol. 41, Cambridge University Press, Cambridge, 2004.
10. R. I. McLachlan and G.R.W. Quispel, Splitting methods, *Acta Numer.* **11** (2002), 341–434.
11. R. I. McLachlan and G. R. W. Quispel, Explicit geometric integration of polynomial vector fields, *BIT* **44** (2004), 515–538.
12. R. I. McLachlan and G. R. W. Quispel, Geometric integrators for ODEs, *J. Phys. A* **39** (2006), 5251–5286.
13. G.R.W. Quispel, Volume-preserving integrators, *Phys. Lett. A* **206**(1–2) (1995), 26–30.
14. G. Rangarajan, Symplectic completion of symplectic jets, *J. Math. Phys.* **37** (1996), 4514–4542.
15. P. Reiterer, C. Lainskcsek, F. Schürrer, C. Letellier, and J. Maquet, A nine-dimensional Lorenz system to study high-dimensional chaos, *J. Phys. A: Math. Gen.* **31** (1998), 7121–7139.
16. Z. J. Shang, Construction of volume-preserving difference schemes for source-free systems via generating functions, *J. Comput. Math.* **12**(3) (1994), 265–272.
17. A. Zanna and H. Z. Munthe-Kaas, Generalized polar decompositions for the approximation of the matrix exponential, *SIAM J. Matrix Anal.* **23**(3) (2002), 840–862.