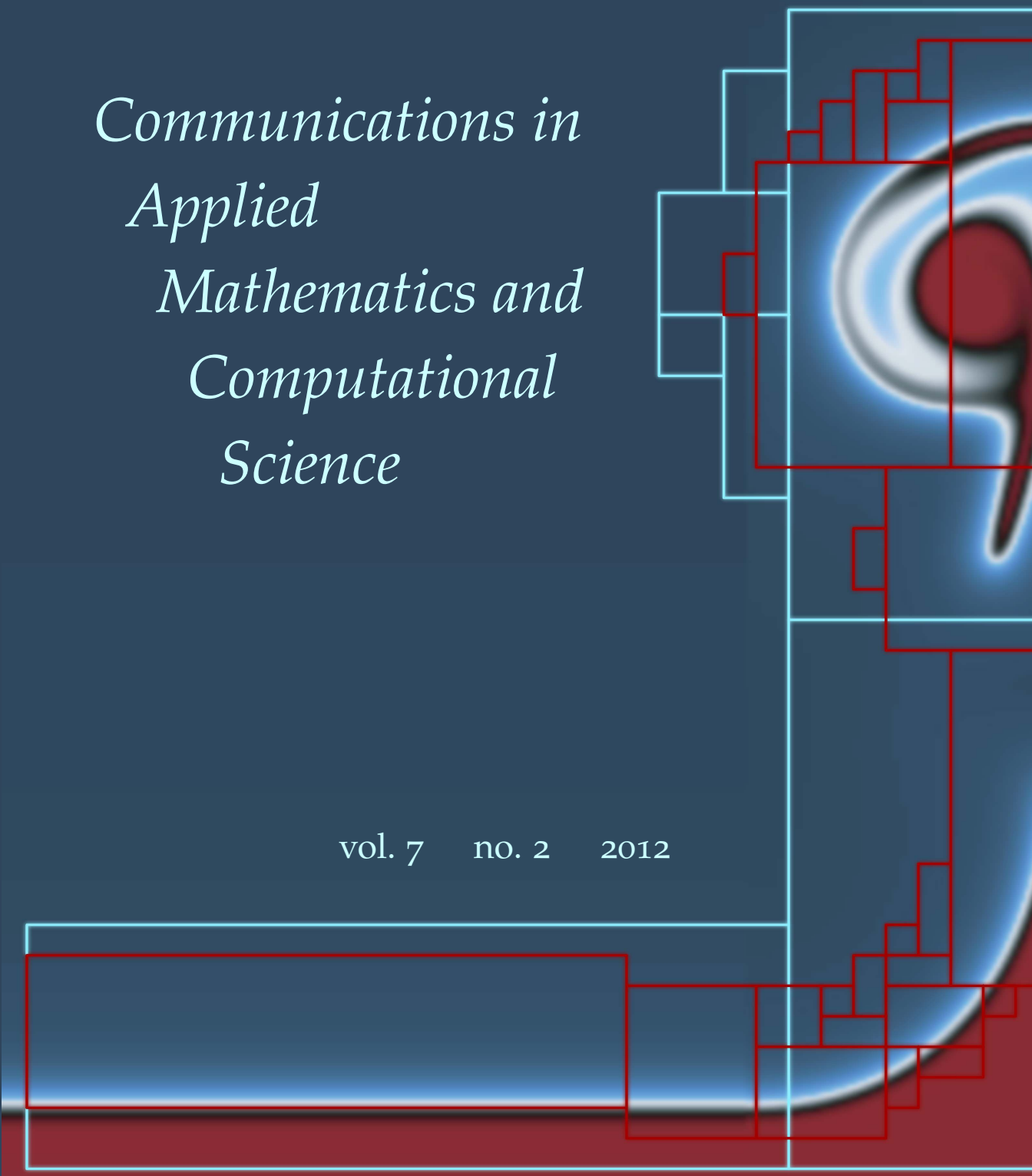
Communications in Applied Mathematics and Computational Science



vol. 7 no. 2 2012

msp

Communications in Applied Mathematics and Computational Science

map.berkeley.edu/camcos

See inside back cover or msp.berkeley.edu/camcos for submission instructions.

The subscription price for 2012 is US $75/year for the electronic version, and $105/year for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94720-3840, USA.

Communications in Applied Mathematics and Computational Science, at Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

CAMCoS peer review and production are managed by EditFLOW™ from Mathematical Sciences Publishers.

PUBLISHED BY
mathematical sciences publishers
http://msp.org/
A NON-PROFIT CORPORATION
Typeset in LaTeX

COMM. APP. MATH. AND COMP. SCI.
Vol. 7, No. 2, 2012
dx.doi.org/10.2140/camcos.2012.7.133

msp

DISCONTINUOUS GALERKIN METHOD WITH THE SPECTRAL DEFERRED CORRECTION TIME-INTEGRATION SCHEME AND A MODIFIED MOMENT LIMITER FOR ADAPTIVE GRIDS

LEANDRO D. GRYNGARTEN, ANDREW SMITH AND SURESH MENON

The discontinuous Galerkin (DG) method is combined with the spectral deferred correction (SDC) time integration approach to solve the fluid dynamic equations. The moment limiter is generalized for nonuniform grids with hanging nodes that result from adaptive mesh refinement. The effect of characteristic, primitive, or conservative decomposition in the limiting stage is studied. In general, primitive variable decomposition is a better option, especially in two and three dimensions. The accuracy-preserving total variation diminishing (AP-TVD) marker for troubled-cell detection, which uses an averaged-derivative basis, is modified to use the Legendre polynomial basis. Given that the latest basis is generally used for DG, the new approach avoids transforming to the averaged-derivative basis, what results in a more efficient technique. Further, a new error estimator is proposed to determine where to refine or coarsen the grid. This estimator is compared against other estimator used in the literature and shows an improved performance. Canonical tests in one, two, and three dimensions are conducted to show the accuracy of the solver.

1. Introduction

The discontinuous Galerkin (DG) method belongs to the finite element (FE) family and uses a piecewise discontinuous space for the test function and the numerical solution [7]. The use of the same function space for the test function and solution defines all Galerkin methods. Usually, the basis to form the space is composed of Legendre [7] or Lagrange [10] polynomials, although other options have been studied in the literature [43]. The discontinuity is localized at the boundary of each element and the coupling between elements is done by computing fluxes as in finite volume (FV) schemes, e.g., using an approximate Riemann solver. This kind of coupling allows DG to formulate each element locally, making the implementation

MSC2010: 35L65, 35L67, 65L06, 65M50, 65M60.
Keywords: discontinuous Galerkin, moment limiter, high-order accuracy, adaptive mesh, troubled-cell detector, spectral deferred correction.

highly parallelizable, *h-p* adaptivity friendly, compatible with complex geometries, and is capable of achieving high-orders of accuracy even with unstructured grids and hanging nodes (e.g., see Remacle et al. [33]). Given that DG is a result of FE and FV, the terms element and cell are generally used indistinctly in this context.

The time integration scheme most widely used has been the ubiquitous 3rd-order TVD Runge–Kutta (RK) method, leading to what is known as the RKDG method [5; 4; 3; 6]. Given that DG has the ability to easily achieve high-order spatial accuracy, some effort to maintain comparable time accuracy has been reported [40]. Under some conditions, especially with higher order derivatives, the time step required for stability of the RKDG method can be very limiting. Recently, Xu and Shu [40] suggested that the spectral deferred correction (SDC) method, derived by Dutt et al. [8], may be an alternative time stepping scheme. It has been shown that SDC can be used in an explicit, semiimplicit, or fully implicit form, and it is easy to extend to high-order accuracy in time [27]. Xia et al. [38] studied a semiimplicit SDC method, in addition to other alternative techniques, to use with the local discontinuous Galerkin (LDG) method. SDC combined with DG (SDC-DG) has not yet been used extensively for practical applications. Grooss and Hesthaven [14] used a semiimplicit SDC to solve the incompressible Navier–Stokes with free-surface flows. Here, we report on new results that demonstrate the potential of SDC-DG with explicit integration and compare it against RKDG. Even though Gottlieb et al. [13] presented RK methods of order higher than 3, these schemes are very difficult to derive, while the extension of SDC to any order is straightforward. In addition, TVD-RK methods of 4th-order or greater require the governing equation to be invariant to time reversal [12; 13]. The Euler equations are invariant to this transformation, but the Navier–Stokes (NS) equations are not. Although we do not use the NS equations in this report, viscous fluxes will be included in future studies. Hence, TVD-RK schemes of 4th-order or higher are not considered here. The possibility of an SDC method with the strong stability preserving (SSP) property was studied by Gottlieb et al. [11] and more extensively by Liu et al. [26]. Note that TVD schemes are SSP schemes that were originally derived using the total variation norm [13], instead of a generic norm. Therefore, in practice the TVD and the SSP properties are equivalent, but TVD could be considered a particular case of SSP. Liu et al. [26] showed that SSP-SDC algorithms can be obtained, but the derivation gets very complicated as the order increases and the CFL coefficient is smaller than for the SSP-RK. Thus, in the current study we use SDC without the SSP property.

The current method also combines the SDC-DG approach with adaptive mesh refinement (AMR) to dynamically and locally refine or coarsen the grid based on an estimation of the numerical error. Issues with the implementation such as hanging nodes, particularly in quadrilateral or hexahedral grids are addressed. The

DG formulation works well with AMR because of its local nature [33] and its performance is demonstrated in this paper.

As with other numerical approaches, it is well known that DG methods may cause nonphysical oscillations close to discontinuities due to the Gibbs phenomenon, especially when higher order schemes are used because of lower numerical dissipation. Therefore, some approach to "limit" this effect is needed. One common technique consists of applying limiters inherited from FV techniques, several of which have been developed in the last two decades. Cockburn and Shu [5] demonstrated a modified *minmod* limiter for the DG method, but it has the disadvantages of dropping the order of accuracy when it is activated and relies on a user-defined parameter to make it total variation bounded (TVB) instead of total variation diminishing (TVD). Qiu and Shu [31] showed that the weighted essentially nonoscillatory (WENO) approach, borrowed from FV, can smooth the un-desired oscillations but increases the size of the stencil and loses the subcell information that DG provides. In a later study, Qiu and Shu [30] used a modified WENO scheme based on Hermite polynomials to reduce the stencil.

Other limiters, such as the moment limiter (ML), originally proposed by Biswas et al. [2] for uniform grids and further improved, e.g., by Krivodonova [22], has also been proposed for DG applications. The ML is generally applied to a Legendre polynomial basis limiting the conservative or the characteristic variables. Yang and Wang [42] modified the ML for unstructured grids for a spectral difference (SD) method, applying it to a polynomial basis based on the averaged derivatives along the cell, instead of estimating the derivatives at the cell center as in [22]. The hierarchical reconstruction (HR) method, introduced by Liu et al. [24], was applied to DG with a WENO-type reconstruction at each hierarchical level [41]. In this approach characteristic decomposition is not used, but rather small overshoots/undershoots appear especially as the order of accuracy is increased [25]. For DG schemes with very high order elements, artificial dissipation to smooth out discontinuities has also been proposed [16; 28; 1]. In this paper the ML as presented in [22] is modified for nonuniform grids with hanging nodes. The ML is usually applied to characteristic variables, which is only consistent in a one-dimensional sense. Therefore, we study the consequences of limiting the conservative, primitive, or characteristic variables to later apply it to multidimensional cases.

Even though good limiters tend to keep the original order of accuracy in smooth regions, they may increase the error slightly [29; 22]. Hence, the application of such limiters within the domain needs to be minimized. This task is carried out by what is usually called a troubled-cell detector, which identifies the cells that may be becoming oscillatory or unstable, and thus require a limiter. Moreover, if the detector is computationally faster than the limiter, the speed of the solver can be

increased by reducing the number of cells where the limiter is applied. In the past, several limiters were adapted as detectors [29], and the ones with best success are the minmod-based TVB limiter [5], the shock-detector by Krivodonova et al. [21] (KXRCF), and the indicator based on Harten's subcell resolution [15]. In [42], the accuracy-preserving TVD (AP-TVD) detector is suggested in an SD frame and compared against the other detectors just mentioned above and was shown to produce better agreement. Therefore, we adapt the AP-TVD to the DG method with some additional modifications, as reported below.

AMR requires an indicator to determine where to refine or coarsen the grid based on an estimated numerical error. This numerical error depends on the scheme, thus error estimators used in FV or finite differences (FD) are not valid here. Considerable research has been invested in estimating the numerical error for the DG method for conservative hyperbolic equations [9], but usually these approaches are computationally expensive and therefore inefficient. Faster, though perhaps less accurate methods have also been derived for DG. Remacle et al. [33] used a simple error estimator based on the jump between elements, which is the same principle as used in the shock-detector KXRCF. Trouble-cell detectors have also been used as error estimators [34; 45]. Zhu et al. [45] compared a few of them and found that KXRCF provided very good results for typical one-dimensional shock problems. In addition, Leicht and Hartmann [23] used the jump between elements to determine the direction for anisotropic refinement. In this study we propose a new estimator which results from a combination of some of these detectors and has better efficiency.

The current paper is organized in the following way. Section 2 introduces the governing equation relevant to this study. Section 3 presents the numerical schemes and algorithm behind the solver. Section 4 includes the test cases that show the success of the method being proposed. Finally, Section 5 summarizes the observations and suggests areas where future work is necessary.

2. Governing equations

The governing equations are the conservation laws written in the general form

$$\begin{cases} \dfrac{\partial \boldsymbol{u}}{\partial t} + \nabla \cdot \boldsymbol{F}(\boldsymbol{u}) = 0 & \text{for } t > 0, \\ \boldsymbol{u} = \boldsymbol{u}_0 & \text{for } t = 0, \end{cases} \tag{1}$$

where $\boldsymbol{u}$ is the solution vector, $\boldsymbol{F}$ is the inviscid flux, and $\boldsymbol{u}_0$ is the initial value. Unless specified otherwise, they correspond to the Euler equation, so that

$$\boldsymbol{u} = (\rho, \rho v_1, \rho v_2, \rho v_3, \rho E_T)^T \tag{2}$$

and

$$\boldsymbol{F}_1 = \begin{pmatrix} \rho v_1 \\ \rho v_1^2 + p \\ \rho v_1 v_2 \\ \rho v_1 v_3 \\ (\rho E_T + p) v_1 \end{pmatrix} \quad \boldsymbol{F}_2 = \begin{pmatrix} \rho v_2 \\ \rho v_1 v_2 \\ \rho v_2^2 + p \\ \rho v_2 v_3 \\ (\rho E_T + p) v_2 \end{pmatrix} \quad \boldsymbol{F}_3 = \begin{pmatrix} \rho v_3 \\ \rho v_1 v_3 \\ \rho v_2 v_3 \\ \rho v_3^2 + p \\ (\rho E_T + p) v_3 \end{pmatrix},$$

where ρ is the density, v_j is the velocity component in the x_j direction, p is the pressure, E_T is the total energy defined as $E_T = e + \sum_{i=1}^{3} \frac{1}{2} v_i^2$ where e is the internal energy. The ideal gas equation of state is $\rho e = p/(\gamma - 1)$ where γ is the specific heat ratio and it is assumed constant. In addition, the speed of sound c is

$$c = \sqrt{\gamma \frac{p}{\rho}}. \tag{3}$$

The state vector $\boldsymbol{u}$ in (2) is given in conservative form. The state vector for the primitive form as used in this study is

$$\boldsymbol{u}_p = (\rho, v_1, v_2, v_3, p)^T. \tag{4}$$

Although in this paper we focus on the Euler equations to show the ability and accuracy of the proposed method, extension to full Navier–Stokes equations are also being evaluated and will be reported in the near future.

3. Numerical method

The DG method is applied to the conservation law described in (1). The domain, Ω, is divided into N nonoverlapping elements:

$$\Omega = \bigcup_{l=1}^{N} \Omega_l. \tag{5}$$

The solution vector $\boldsymbol{u}$ is approximated per element by $\boldsymbol{U}_l$, defined by the basis ϕ:

$$\boldsymbol{U}_l = \sum_{i=0}^{p} \phi_i \boldsymbol{c}_{i,l}, \tag{6}$$

where p defines the order of the finite element and $\boldsymbol{c}_{i,l}$ is the weight corresponding to each element of the basis ϕ. After multiplying by the test function, which is equal to the basis ϕ, and integrating by parts we arrive at the weak form of the DG method [7]:

$$\begin{cases} \displaystyle\int_{\Omega_l} \phi \frac{\partial \boldsymbol{U}_l}{\partial t} dV - \int_{\Omega_l} \nabla\phi \cdot \boldsymbol{F}(\boldsymbol{U}_l)\, dV + \int_{\partial\Omega_l} \phi\, \widehat{F}(\boldsymbol{U}^-, \boldsymbol{U}^+)\, dS = 0 & \text{for } t > 0, \\ \displaystyle\int_{\Omega_l} \phi \boldsymbol{U}_l\, dV = \int_{\Omega_l} \phi \boldsymbol{u}_0\, dV & \text{for } t = 0, \end{cases} \tag{7}$$

with appropriate boundary conditions. Here, $\widehat{F}$ is a numerical flux normal to the boundary of the element and needs to be properly defined given that it is computed at the face of the elements, which may be discontinuous. $\boldsymbol{U}^-$ is the value of $\boldsymbol{U}_l$ according to the current element l at the face and $\boldsymbol{U}^+$ is the value of $\boldsymbol{U}_m$ at the face based on the neighboring element m.

The numerical flux should be an exact or approximate Riemann solver. Here the local Lax–Friedrichs flux is used as it is known to provide good results and is simple to compute [5]. In this study, the spatial integration in (7) is done with a full quadrature rule using Lobatto points [20; 5; 4; 3; 6].

3.1. ***Time integration.*** Time integration is conducted explicitly using the Runge–Kutta (RK) method or the spectral deferred correction (SDC) method. Both approaches treat the governing equations as a system of ordinary differential equations (ODE):

$$\frac{d\boldsymbol{u}}{dt} = G(t, \boldsymbol{u}), \tag{8}$$

where $G(t, \boldsymbol{u}) = -\nabla \cdot \boldsymbol{F}(\boldsymbol{u})$. In this study, unless specified otherwise, for elements of polynomial order p a time integration of order $p+1$ is used. Unless specified otherwise, the time step is given by

$$\Delta t = \min_{l=1\ldots N} \left[\frac{C}{2p_l + 1} \cdot \min_{i=1\ldots n_d} \left(\frac{\Delta x_{i,l}}{v_{i,l} + c_l} \right) \right], \tag{9}$$

where n_d is the number of dimensions and C is a constant. The flow velocity $v_{i,l}$ and the speed of sound c are considered at the centroid of element l. We use $C = 0.5$ unless specified otherwise. It is usually replaced by 1.0 (see [7]), however, in this study we choose 0.5 to be more conservative. Maximum stable time-step size for RKDG has been shown elsewhere [7]. Stability limits for SDC-DG have not been studied in the literature, at least to the author's knowledge. Equation (9) turns out to provide a stable condition for the tests presented in this study also for SDC-DG when the order in time is equal to $p+1$.

3.1.1. *The Runge–Kutta method.* The Runge–Kutta method is a well known family of schemes. The current study used the total variation diminishing RK (TVD-RK) of second and third orders [35], which can be summarized in three steps:

Step 1:

$$u^0 = u_n. \tag{10}$$

Step 2:

$$u^i = \sum_{l=0}^{i-1} \alpha_{il} w^{il}, \quad w^{il} = u^l + \frac{\beta_{il}}{\alpha_{il}} \Delta t G\left(u^l, t + \Delta t \cdot d_l\right), \quad \text{for } i = 1, \ldots, K. \tag{11}$$

Step 3:

$$u_{n+1} = u^K, \tag{12}$$

where the superindexes of u determine intermediate steps between u_n and u_{n+1}. The parameters are given in Table 1. A good property for these 2nd- and 3rd-order TVD schemes is that if for some seminorm $|\cdot|$, we have that $|w^{il}| \le |u^l|$, then $|u_{n+1}| \le |u_n|$.

Order	α_{il}	β_{il}	d_l
2	1	1	0
	1/2 1/2	0 1/2	1
3	1	1	0
	3/4 1/4	0 1/4	1
	1/3 0 2/3	0 0 2/3	1/2

Table 1. Parameters for TVD-RK of order 2 and 3.

3.1.2. *The spectral deferred correction method.* Although details of this method are given elsewhere [8; 27; 26; 40], we include the main algorithm for completeness. The scheme is based on first-order explicit integration of substeps and iterative correction [8]. For stiff problems, the scheme can be varied with a more implicit character, but only the explicit method is addressed here. Each time step $[t_n, t_{n+1}]$ is divided into J substeps: $t_n = t_{n,0} < t_{n,1} < \cdots < t_{n,m} < t_{n,m+1} < \cdots < t_{n,J} = t_{n+1}$. These points are chosen as quadrature points (Lobatto points in the current study). This approach makes the scheme more stable because it avoids a uniform distribution and leads to the spectral characteristic of the scheme [8]. This property is important to stabilize higher orders. Initially, the governing equations are integrated with a first-order explicit integration from t_n to t_{n+1} using $t_{n,m}$ points:

$$u^1_{n,m+1} = u^1_{n,m} + \Delta t_{n,m} G\left(t_{n,m}, u^1_{n,m}\right) \quad \text{for } m = 0, \ldots, J-1, \tag{13}$$

where $u^1_{n,0} = u_n$ and $\Delta t_{n,m} = t_{n,m+1} - t_{n,m}$.

Now K iterations are computed for $k = 1, \ldots, K$ and $m = 0, \ldots, J-1$ (m being the inner loop):

$$u^{k+1}_{n,m+1} = u^{k+1}_{n,m} + \theta \Delta t_{n,m} \left(G(t_{n,m}, u^{k+1}_{n,m}) - G(t_{n,m}, u^k_{n,m})\right) + I^{m+1}_m \left(G(t_{n,m}, u_{n,m})\right), \tag{14}$$

where $0 \le \theta \le 1$ and $I^{m+1}_m(G(t_{n,m}, u_{n,m}))$ is the integral of the interpolating polynomial along the quadrature points:

$$I^{m+1}_m(G(t_{n,m}, u_{n,m})) = \int_{t_{n,m}}^{t_{n,m+1}} G(\tau, u(\tau))\, d\tau. \tag{15}$$

Finally, $u_{n+1} = u_{n,J}^{K+1}$. Here, we use $\theta = 1$ as in the original study [8] and $K = J - 1$. For the cases studied in this report, we observed that neglecting the second term on the right-hand side, i.e., using $\theta = 0$, provides similar results but with greater numerical error.

3.2. ***The basis.*** Several options can be used to form the finite element space. Our basis ϕ is built on the Legendre polynomials P_i, which leads to an orthogonal, hierarchical, polynomial basis — an advantage in comparison with computationally more expensive functions (e.g., trigonometric or exponential). Another numerical advantage is an advantage in comparison with computationally more expensive functions the lower condition number of the Vandermonde matrix, which transforms from modal space to nodal space. The mass matrix is diagonal when the basis is orthogonal and the Jacobian is constant inside the element; indeed, if the basis is correctly normalized and the Jacobian is constant, the mass matrix is just the identity matrix times the Jacobian.

The normalized Legendre polynomials are given by

$$\phi_i = P_i \sqrt{\frac{2i+1}{2}} \quad \text{for } i = 0, \ldots, p, \tag{16}$$

and they are orthonormal: $\int_{\Omega_l} \phi_i \phi_j dV = \delta_{ij}$, where δ_{ij} is Dirac's delta function.

For quadrilateral and hexahedral elements the basis can easily be generated from the 1D basis by applying a tensor product. Thus, in 2D we have

$$\phi_{ij}(\xi, \eta) = \phi_i(\xi)\ \phi_j(\eta), \tag{17}$$

and in 3D

$$\phi_{ijk}(\xi, \eta, \zeta) = \phi_i(\xi)\ \phi_j(\eta)\ \phi_k(\zeta). \tag{18}$$

3.3. ***Adaptive mesh refinement.*** The solver relies on a tree to handle the hierarchical structure of the grid adaptations. The initial grid is composed of *root* cells, corresponding to the lowest level. Each cell can have children. A cell that does not have children is called a *leaf* cell. If a root cell does not have children it is also tagged as a leaf cell. The root cells correspond to level 1 and the maximum level is given by $\ell_{\max}$, which may depend on the problem. Each face of every cell has to be connected to a neighbor or to a boundary element. A cell can connect to a neighbor at the same level or at a lower level, but never at a higher level. In addition, there is a ghost tree to handle the ghost cells for interprocessor communications. Details about the tree structures are given in [19; 17].

When a cell is marked for refinement it is split into two, four, or eight, in dimension 1, 2, or 3. The variables from the parent are projected onto the children with an identity projection. On the other hand, when all the children of a cell are marked for coarsening, the variables from the children are projected onto the parent cell

with a least-squares projection. Note that if the order p is kept constant, no data is lost when refinement is done; however, data is lost when coarsening is done.

Cells are marked for adaptation based on an error estimator. The error ϵ_l of cell l is then normalized by the maximum error $\epsilon_{\max}$ found in the whole domain. Then, a logarithmic scale is applied as in [9]. The current hierarchical level in the tree for cell l is $\ell(l)$. A target level ℓ_t is estimated as

$$\ell_t = \max\left(1, \ell_{\max} - \mathrm{INT}\left(\log(\epsilon_{\max}/\epsilon_j)/\log d\right)\right) \tag{19}$$

where d is a parameter that determines the sensibility of the refinement, the larger its value, the more refinement will be done. Even though the accuracy is expected to increase as d is raised, the computational cost will be higher too. The default value adopted here is $d = 10$ as in [9], which is a good balance between computational cost and accuracy. If ℓ_t is greater than $\ell(l)$ then cell l is marked for refinement. If ℓ_t is lower than $\ell(l)$ then cell l is merged for coarsening. Note that for coarsening to actually be feasible, all the children have to be marked for coarsening.

The level difference between neighboring cells is not allowed to be larger than 1. For example, suppose cells 1 and 2 are neighbors, with $\ell(1) = 3$ and $\ell(2) = 4$. If cell 2 is marked for refinement, then cell 1 will be marked for refinement also.

For the sake of simplicity and computational speed, a simple error estimator is used here. More accurate approaches are slower and may increase the overhead, making the adaptivity too costly.

Zhu et al. [45] compared a few different shock-detectors as estimators for refinement, and concluded that the most efficient based on their 1D discontinuous problems was the KXRCF [21]:

$$\epsilon_{A,l} = \frac{\left|\int_{\delta\Omega_l^-} (U^- - U^+)\, dS\right|}{h_l^{\frac{p+1}{2}} \int_{\delta\Omega_j^-} dS\, \|U_l\|}, \tag{20}$$

where U is some relevant variable, $\delta\Omega_l^-$ is the element boundary where the velocity is going into the element, h_l is the radius of the circle circumscribed to the element l, and the norm is based on an element average. In [33] the following error estimator was used for element l:

$$\epsilon_{B,l} = \int_{\Omega_l} |U^- - U^+|\, dS. \tag{21}$$

Here, we combine the best of (20) and (21) to obtain

$$\epsilon_{C,l} = \frac{\int_{\delta\Omega_l} |U^- - U^+| dS}{\int_{\delta\Omega_l} dS.} \tag{22}$$

In the current implementation the error is already normalized by the maximum error in the domain — see (19) — and so the additional normalization needed in (20) is not required. For the Euler equation two error estimators based on the density and the total energy are used. Below we refer to the estimators in (20)–(22) as KXRCF, JUMP1, and JUMP2, respectively. The difference between JUMP1 and JUMP2 can only be observed in 2 and 3 dimensions, so for the 1D cases JUMP1 is not used.

3.4. ***Moment-limiter for nonuniform grids.*** The limiting strategy of the ML is shown below for 1D. However, for completeness, the 2D and 3D extensions are discussed in the Appendix. In Section 3.4.2, we extend the original ML to nonuniform grids for 1D, but its extension to higher dimensions is trivial, except for when a neighbor is split due to refinement, in which case the average of the two is used, and when the neighbor is coarser, in which case a virtual refinement of the neighbor is done. This last step has no analytical complexity, but its implementation may not be trivial. Figure 1 shows the stencil used for limiting purposes when coarser, finer, or equal level neighbors are present. For the neighbor on the right-hand side, a virtual refinement was created, similar to the idea of partial neighboring cells in [41]. The neighbors on the top are virtually merged.

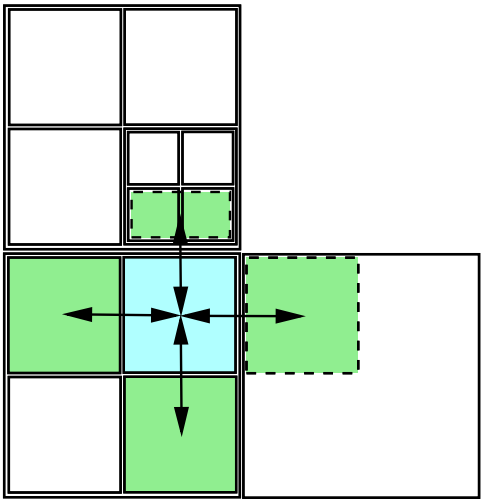


Figure 1. Example of a 2D stencil used for limiting when coarser, finer, or equal-level neighbors are present.

3.4.1. *The moment-limiter concept.* The idea is to limit the i-th derivative in x of U_l in the following way:

$$\frac{\partial^i \tilde{U}_l}{\partial x^i} = \text{minmod}\left(\frac{\partial^i U_l}{\partial x^i}, \beta_i D_i^+, \beta_i D_i^-\right) \tag{23}$$

where

$$\operatorname{minmod}(a,b,c) = \begin{cases} \operatorname{sgn} a \ \min(|a|,|b|,|c|) & \text{if } \operatorname{sgn} a = \operatorname{sgn} b = \operatorname{sgn} c, \\ 0 & \text{otherwise,} \end{cases} \tag{24}$$

and $\tilde{U}_l$ is the solution U_l after the limiter is applied. $D_i^{+/-}$ is an estimation of the i-th derivative based on one-sided differences:

$$D_i^+ = \frac{\dfrac{\partial^{i-1} U_{l+1}}{\partial x^{i-1}} - \dfrac{\partial^{i-1} U_l}{\partial x^{i-1}}}{\bar{x}_{l+1} - \bar{x}_l}, \quad D_i^- = \frac{\dfrac{\partial^{i-1} U_l}{\partial x^{i-1}} - \dfrac{\partial^{i-1} U_{l-1}}{\partial x^{i-1}}}{\bar{x}_l - \bar{x}_{l-1}}, \tag{25}$$

where $\bar{x}_l$ is the location of the centroid of element l, and β_i is a parameter to control the sensibility of the limiter. If there is a boundary condition against one of the faces of the element, then that side is neglected in (23).

In the literature (e.g., [22]) it is recommended to apply limiting to the characteristic variables when a system of equations is being solved. This means replacing (23) by

$$\left(\boldsymbol{L} \frac{\partial^i \tilde{\boldsymbol{U}}_l}{\partial x^i} \right)_k = \operatorname{minmod}\left(\left(\boldsymbol{L} \frac{\partial^i \boldsymbol{U}_l}{\partial x^i} \right)_k, \beta_i (\boldsymbol{L} \boldsymbol{D}_i^+)_k, \beta_i (\boldsymbol{L} \boldsymbol{D}_i^-)_k \right), \tag{26}$$

where $\boldsymbol{L}$ is a matrix composed by the left eigenvectors of the Jacobian $\partial \boldsymbol{F} / \partial \boldsymbol{u}$, and the subindex k refers the k-th characteristic variable. Each characteristic variable is limited individually and this means that if a variable in a given element is not limited the others can still be limited. If limiting is applied in an element the resulted characteristic variables have to be converted back to the conservative variables, multiplying the characteristic variables by the inverse of $\boldsymbol{L}$, which is composed by the right eigenvectors of $\partial \boldsymbol{F} / \partial \boldsymbol{u}$. In addition, if one wants to use primitive variables for this stage, $\boldsymbol{L}$ should be replaced by the Jacobian, $\partial \boldsymbol{u}_p / \partial \boldsymbol{u}$, where $\boldsymbol{u}_p$ is the state vector in primitive variables.

The algorithm to apply the limiter is the following:

(1) Apply (23) for $i = p$ to every element. If

$$\frac{\partial^i \tilde{U}_l}{\partial x^i} = \frac{\partial^i U_l}{\partial x^i}, \tag{27}$$

then mark the element as not needing limiting anymore.

(2) Apply (23) for $i = p - 1$ to every element that still needs to be limited.

(3) Continue for $i = p - 2, \ldots, 1$ or until no element requires limiting.

Note that only the derivatives are modified, not the mean value; thus the limiter does not violate the conservation property.

As in [22], we also add the following steps at the end of the limiting procedure for each element to avoid nonphysical values:

(1) If any integration point has a nonphysical state (e.g., negative pressure), make all the quadratic and higher-order moments equal to zero and go to step 2, otherwise the procedure is completed.

(2) If any integration point still has a nonphysical state (e.g., negative pressure), make all the linear moments equal to zero. This makes the solution piecewise constant.

Obviously, when these steps are applied the accuracy is forced to drop locally. Nonetheless, this is not needed often.

3.4.2. *The ML using a Legendre basis.* The $(i-1)$-th derivative with respect to x of $\boldsymbol{U}_l$, given in (6), can be expressed as

$$\frac{\partial^{i-1}\boldsymbol{U}_l}{\partial x^{i-1}} = \left(\frac{2}{\Delta x_l}\right)^{i-1}\left[\sqrt{\frac{2i-1}{2}}\,(2i-3)!!\,\boldsymbol{c}_{l,i-1} + \frac{\partial^{i-1}}{\partial \xi^{i-1}}\sum_{k=i}^{p}\boldsymbol{c}_{l,k}\phi_k(\xi)\right] \tag{28}$$

and the i-th derivative in x of (6) can be expressed as

$$\frac{\partial^{i}\boldsymbol{U}_l}{\partial x^{i}} = \left(\frac{2}{\Delta x_l}\right)^{i}\left[\sqrt{\frac{2i+1}{2}}\,(2i-1)!!\,\boldsymbol{c}_{l,i} + \frac{\partial^{i}}{\partial \xi^{i}}\sum_{k=i+1}^{p}\boldsymbol{c}_{l,k}\phi_k(\xi)\right], \tag{29}$$

where Δx_l is the length of element l.

At the same time, the i-th derivative could be estimated from the forward or backward differences of $\partial^{i-1}\boldsymbol{U}_l/\partial x^{i-1}$:

$$\frac{\partial^{i}\boldsymbol{U}_l}{\partial x^{i}} = \left(\frac{\partial^{i-1}U_{l+1}}{\partial x^{i-1}} - \frac{\partial^{i-1}U_l}{\partial x^{i-1}}\right)\frac{2}{\Delta x_{l+1}+\Delta x_l}, \tag{30}$$

$$\frac{\partial^{i}\boldsymbol{U}_l}{\partial x^{i}} = \left(\frac{\partial^{i-1}U_{l}}{\partial x^{i-1}} - \frac{\partial^{i-1}U_{l-1}}{\partial x^{i-1}}\right)\frac{2}{\Delta x_{l}+\Delta x_{l-1}}. \tag{31}$$

Therefore, ignoring higher order derivatives we obtain

$$\boldsymbol{c}_{l,i} = \frac{2\vartheta_+}{1+\vartheta_+}\sqrt{\frac{2i-1}{2i+1}}\frac{1}{2(2i-1)}\left(\vartheta_+^{i-1}\boldsymbol{c}_{l+1,i-1} - \boldsymbol{c}_{l,i-1}\right), \tag{32}$$

$$\boldsymbol{c}_{l,i} = \frac{2\vartheta_-}{1+\vartheta_-}\sqrt{\frac{2i-1}{2i+1}}\frac{1}{2(2i-1)}\left(\boldsymbol{c}_{l,i-1} - \vartheta_-^{i-1}\boldsymbol{c}_{l-1,i-1}\right), \tag{33}$$

where $\vartheta_- = \Delta x_l/\Delta x_{l-1}$ and $\vartheta_+ = \Delta x_l/\Delta x_{l+1}$. Note that if the grid is uniform $\vartheta_- = 1$, $\vartheta_+ = 1$, and the derived equations converge to the solution in [22]. Thus, the difference between the current derivation and the one in [22] starts in (30) and (31), where we do not assume a constant Δx.

One could apply the limiter as

$$\tilde{\boldsymbol{c}}_{l,i} = \text{minmod}\left(\boldsymbol{c}_{l,i},\, \Delta_i^+,\, \Delta_i^-\right), \tag{34}$$

where Δ^+ and Δ^- are the right-hand sides of (32) and (33). However, to make the limiter less numerically diffusive, [22] uses an expression equivalent to

$$\tilde{\boldsymbol{c}}_{l,i} = \text{minmod}\left(\boldsymbol{c}_{l,i},\, 2\,(2i-1)\,\Delta_i^+,\, 2\,(2i-1)\,\Delta_i^-\right). \tag{35}$$

This equation should be the actual implementation of what (23) represents. The same procedure is easily extended to 2D and 3D. Note that this formulation is equivalent to what was presented in [22] except for the generalization for nonuniform grids and how to handle neighbors of different adaptive level.

3.5. ***Troubled-cell detector.*** The detector presented in this study is a modification of the AP-TVD detector presented in [42] for a spectral difference (SD) scheme. The adapted technique consists of two steps:

1. For each cell l compute

$$\bar{\boldsymbol{U}}_{\max,l} = \max\left(\bar{\boldsymbol{U}}_{l-1}, \bar{\boldsymbol{U}}_l, \bar{\boldsymbol{U}}_{l+1}\right), \tag{36}$$

$$\bar{\boldsymbol{U}}_{\min,l} = \min\left(\bar{\boldsymbol{U}}_{l-1}, \bar{\boldsymbol{U}}_l, \bar{\boldsymbol{U}}_{l+1}\right), \tag{37}$$

   where $\bar{\boldsymbol{U}}_l$ indicates the average of $\boldsymbol{U}$ in cell l. If for any node i in element l we have $\boldsymbol{U}_{i,l} > 1.001\,\bar{\boldsymbol{U}}_{\max,l}$ or $\bar{U}_{i,l} < 0.999\,\bar{\boldsymbol{U}}_{\min,l}$, then proceed to step 2; else the element is not marked.

2. For each dimension j the idea is to compute

$$\frac{\partial^2 \tilde{\boldsymbol{U}}_l}{\partial x_j^2} = \text{minmod}\left(\frac{\partial^2 \boldsymbol{U}_l}{\partial x_j^2},\ \beta\frac{\dfrac{\partial \boldsymbol{U}_{l+1}}{\partial x_j} - \dfrac{\partial \boldsymbol{U}_l}{\partial x_j}}{x_{l+1} - x_j},\ \beta\frac{\dfrac{\partial \boldsymbol{U}_l}{\partial x_j} - \dfrac{\partial \boldsymbol{U}_{l-1}}{\partial x_j}}{x_l - x_{l-1}}\right). \tag{38}$$

   The derivatives are estimated from the Legendre polynomials as in 3.4.2, so the implementation of (38) is

$$\tilde{\boldsymbol{c}}_{l,2} = \text{minmod}\left(\boldsymbol{c}_{l,2},\ \varrho\vartheta_+\frac{\vartheta_+\boldsymbol{c}_{l+1,1} - \boldsymbol{c}_{l,1}}{1+\vartheta_+},\ \varrho\vartheta_-\frac{\boldsymbol{c}_{l,1} - \boldsymbol{c}_{l-1,1}\vartheta_-}{1+\vartheta_-}\right) \tag{39}$$

   where $\varrho = 2\sqrt{3/5}$. If $\tilde{\boldsymbol{c}}_{2,l} \neq \boldsymbol{c}_{2,l}$ the cell is marked for limiting. According to [42]; β is a parameter between 1 and 2, the higher its value the less dissipative the scheme will be. We use $\beta = 2$ to make it consistent with the ML used here.

There are two main differences in the current limiter with respect to the AP-TVD developed earlier [42]. The first one is in step 1 where every node in the element is tested, while in [42] only the nodes at element boundaries are checked. The second difference is in the way the derivatives are estimated in step 2, averaged-derivatives were used in [42], while here we suggest using estimations of the derivatives, as

done in the ML based on Legendre polynomials to avoid computing the averaged-derivatives. Thus, we call the current detector the moment-based AP-TVD or MB-AP-TVD.

Usually, the detection is done based on the conservative variables, instead of transforming to characteristic or primitive variables, in order to keep this stage as computationally cheap as possible.

4. Results and discussion

Various test cases used in past studies are used to establish the capability of this new numerical algorithm. In addition, we use some 2D and 3D cases to demonstrate the potential of the method for more complex problems. The details of the test cases and the rationals for them are summarized in Table 2.

Test case	Purpose
Order of convergence – linear	Order of accuracy in space with a smooth linear problem.
Order of convergence – nonlinear	Order of accuracy in space with a smooth nonlinear problem.
Accuracy in time	Order of accuracy in time with a smooth solution using RK and SDC.
Advection of mixed pulses	Order of accuracy with a nonsmooth solution with and without detector.
High order for a smooth and nonsmooth solution	Order of accuracy with a localized discontinuity.
Sod's problem	Limiting variables, with and without detector, and with and without AMR.
Lax's problem	Limiting variables, with and without detector, and with and without AMR.
Blast waves	Limiting variables, with and without detector, and with and without AMR.
Shock-entropy waves interaction	Limiting variables, with and without detector, and with and without AMR.
2D convection	Detector and AMR in 2D.
Double-Mach reflection	Example in 2D.
Vortex convection	Smooth example in 2D.
Shock-vortex interaction	Example in 2D.
Spherical shock test	Multidimensional symmetry (3D).

Table 2. Summary of test cases.

4.1. ***Order of convergence – linear.*** The order of convergence is studied with the one-dimensional convection equation because the exact solution is known:

$$\frac{\partial u(x,t)}{\partial t} + c\frac{\partial u(x,t)}{\partial x} = 0, \qquad (40)$$
$$u(x, t=0) = \sin x,$$

where u is a passive scalar and c is the constant convection velocity equal to 1. The exact solution is $u(x,t) = \sin(x - ct)$. The domain has a length of 2π and periodic boundary conditions. The number of cells N and the polynomial order p are varied in this study.

This case is run without detector to show the effect of the limiters on smooth solutions. Moreover, given that the governing equation is not a system of equations, no characteristic decomposition is needed.

The time integration schemes used here are the SDC and the TVD-RK of 3rd order with a time step of 10^{-5}. This gives an error in time of the order of 10^{-15}, which is negligible with respect to the spatial error and of the order of the round-off error. The L_∞ error, e_{L_∞}, at $t = 2$ is computed at the centroid of the element and with respect to the exact solution, i.e.,

$$e_{L_\infty} = \max_{l=1,\ldots,N} |U_l(\bar{x}_l, t) - u(\bar{x}_l, t)| \qquad (41)$$

where $\bar{x}_l$ is the centroid of element l. The L_∞ error in the plots is normalized by the case with the largest error. As shown in Figure 2, elements of order p lead to an order of accuracy of $p+1$, as the literature predicts [7]. Although the solutions with limiter have the same order of accuracy, they have a greater error. Therefore, the limiter should not be used unless really needed. The curves in Figure 2 get flattened out for very low errors (around $O(10^{-11})$) due to accumulated round-off error.

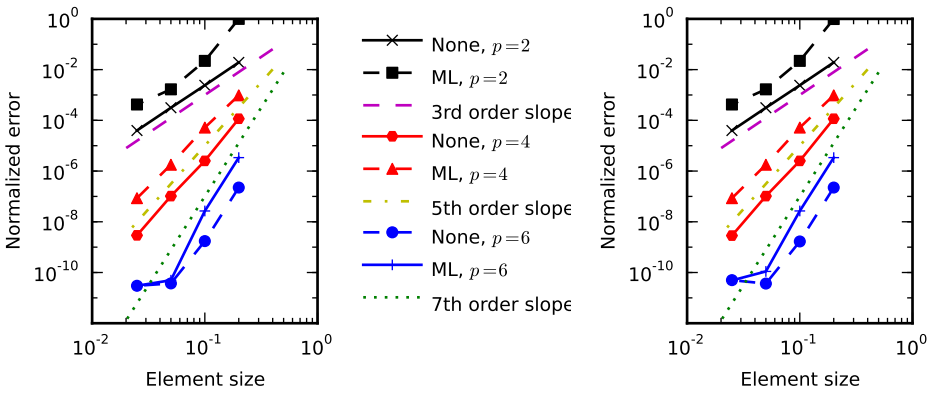


Figure 2. Grid convergence for different orders when a smooth solution is convected, for the 3rd-order TVD-RK (left) and the 3rd-order SDC (right).

4.2. ***Order of convergence – nonlinear.*** Now the order of convergence in space is studied with the one-dimensional burger equation as in [7]:

$$\frac{\partial u(x,t)}{\partial t} + \frac{\partial (u(x,t)^2/2)}{\partial x} = 0, \tag{42}$$

$$u(x, t=0) = \tfrac{1}{4} + \tfrac{1}{2}\sin(\pi(2x-1)),$$

where u is the velocity. The domain has a unit length and periodic boundary conditions. The number of cells N and the polynomial order p are varied in this study. The exact solution is estimated with $N = 2048$, $p = 2$, 3rd-order TVD-RK, and without limiter. The problem is solved until $t = 0.05$, when the solution is still smooth.

This case is run without detector to show the effect of the limiters on smooth solutions.

The time integration schemes used here are the SDC and the TVD-RK of 3rd order with a time step of 10^{-5}. This gives an error in time of the order of 10^{-15}, which is negligible with respect to the spatial error and of the order of the round-off error. The L_∞ error, e_{L_∞}, at $t = 0.05$ is computed at the centroid of the element and with respect to the estimated exact solution as in the previous case. As shown in Figure 3, the order of accuracy in space matches closely with what the theory predicted even for a nonlinear problem. Even though the solutions with limiter have the same order of accuracy, they have a greater error. Therefore, the limiter should not be used if it is not really needed. The curves in Figure 3 get flattened out for very low errors due to accumulated round-off error. In conclusion, the observations for the nonlinear case are very similar to the linear case above.

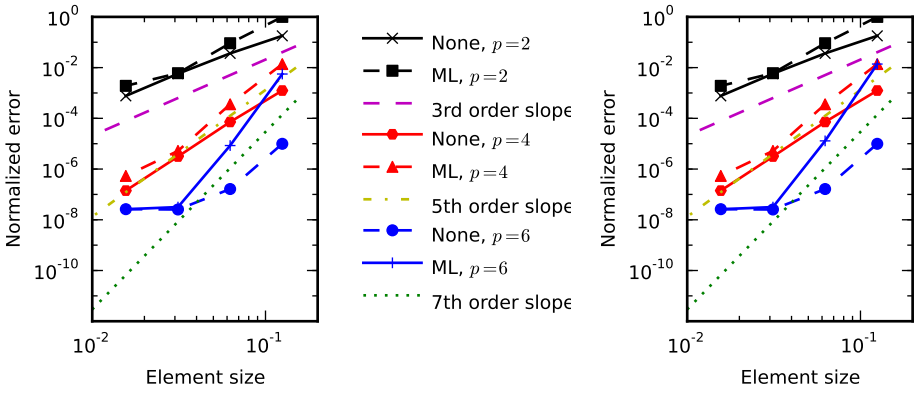


Figure 3. Grid convergence for different orders with a smooth nonlinear problem, for the 3rd-order TVD-RK (left) and the 3rd-order SDC (right).

4.3. ***Accuracy in time.*** The time integration is studied with the equation

$$\frac{\partial u(x,t)}{\partial t} + 0\frac{\partial u(x,t)}{\partial x} = u(x,t), \tag{43}$$
$$u(x,0) = 1,$$

which has the exact solution $u(x,t) = e^t$. The convection velocity is 0 so that the truncation error in space is zero and the truncation error in time can be studied by itself. A 1D domain of unit length, periodic boundaries and 100 elements is used. The time integration is performed with the TVD-RK and SDC methods for different order. The L_∞ error is computed at $t = 6.28$ for different number of time steps and shown in Table 3 along with the order of accuracy. The results are also shown in Figure 4 for a more clear appreciation. The order of accuracy for N_i elements (knowing that $N_i = 2N_{i-1}$) is computed as

$$\frac{\log(e_i/e_{i-1})}{\log(0.5)}. \tag{44}$$

The fact that the computed order approaches the order of the scheme verifies the proper implementation of the temporal integration. Also, note that at equal theoretical order, SDC results to be more accurate while they have very similar order of accuracy.

Number of time steps	2nd order TVD-RK e_{L_∞}	order	3rd order TVD-RK e_{L_∞}	order	2nd order SDC e_{L_∞}	order
8	$1.6537 \cdot 10^{2}$	—	$3.5302 \cdot 10^{1}$	—	$1.6537 \cdot 10^{2}$	—
16	$6.0804 \cdot 10^{1}$	1.4435	$6.1493 \cdot 10^{0}$	2.5213	$6.0804 \cdot 10^{1}$	1.4435
32	$1.8276 \cdot 10^{1}$	1.7342	$9.0205 \cdot 10^{-1}$	2.7691	$1.8276 \cdot 10^{1}$	1.7342
64	$4.9757 \cdot 10^{0}$	1.8770	$1.2200 \cdot 10^{-1}$	2.8863	$4.9757 \cdot 10^{0}$	1.8770
128	$1.2948 \cdot 10^{0}$	1.9422	$1.5861 \cdot 10^{-2}$	2.9434	$1.2948 \cdot 10^{0}$	1.9422
256	$3.2999 \cdot 10^{-1}$	1.9722	$2.0219 \cdot 10^{-3}$	2.9717	$3.2999 \cdot 10^{-1}$	1.9722
Number of time steps	3rd order SDC e_{L_∞}	order	4th order SDC e_{L_∞}	order	5th order SDC e_{L_∞}	order
8	$1.9510 \cdot 10^{1}$	—	$1.2840 \cdot 10^{0}$	—	$7.2084 \cdot 10^{-2}$	—
16	$2.8648 \cdot 10^{0}$	2.7677	$9.2132 \cdot 10^{-2}$	3.8007	$2.4229 \cdot 10^{-3}$	4.8949
32	$3.7984 \cdot 10^{-1}$	2.9150	$6.0812 \cdot 10^{-3}$	3.9213	$7.7303 \cdot 10^{-5}$	4.9700
64	$4.8588 \cdot 10^{-2}$	2.9667	$3.8890 \cdot 10^{-4}$	3.9669	$2.4288 \cdot 10^{-6}$	4.9922
128	$6.1332 \cdot 10^{-3}$	2.9859	$2.4556 \cdot 10^{-5}$	3.9852	$7.5987 \cdot 10^{-8}$	4.9984
256	$7.7005 \cdot 10^{-4}$	2.9936	$1.5422 \cdot 10^{-6}$	3.9931	$2.3588 \cdot 10^{-9}$	5.0096

Table 3. Error e_{L_∞} for TVD-RK (2nd and 3rd order) and for SDC (2nd to 5th order).

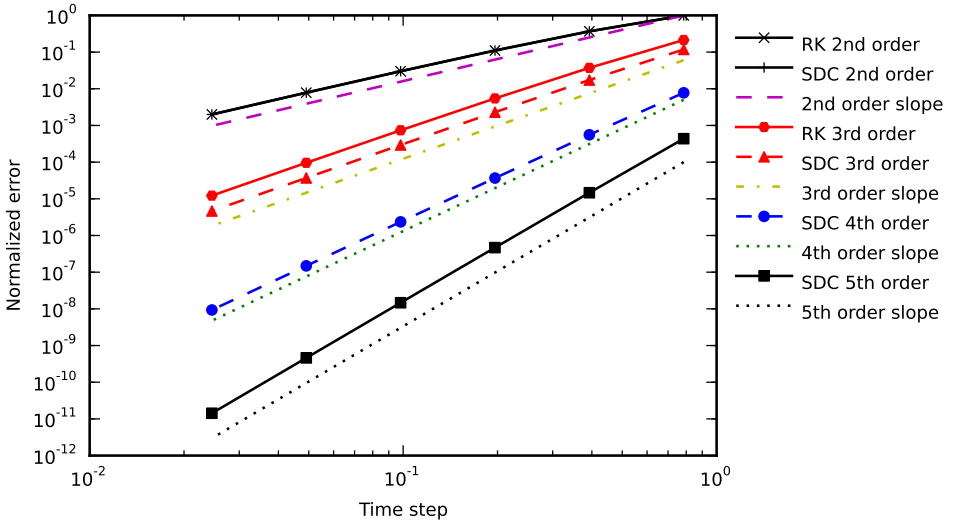


Figure 4. Normalized error for TVD-RK and SDC.

Figure 5 shows the CPU time against the e_{L_∞} obtained for different orders and schemes. The CPU time is normalized by the fastest case. The curves closer to the bottom left corner represent a more efficient scheme. For the same order, TVD-RK is more efficient than SDC. At the same time, the efficiency is increased with the order. For instance for this case 5th order SDC is more efficient than 3rd order TVD-RK.

In conclusion, the advantage of SDC with respect to RK is that the extension to higher orders is trivial. One could argue that SDC does not have the TVD property

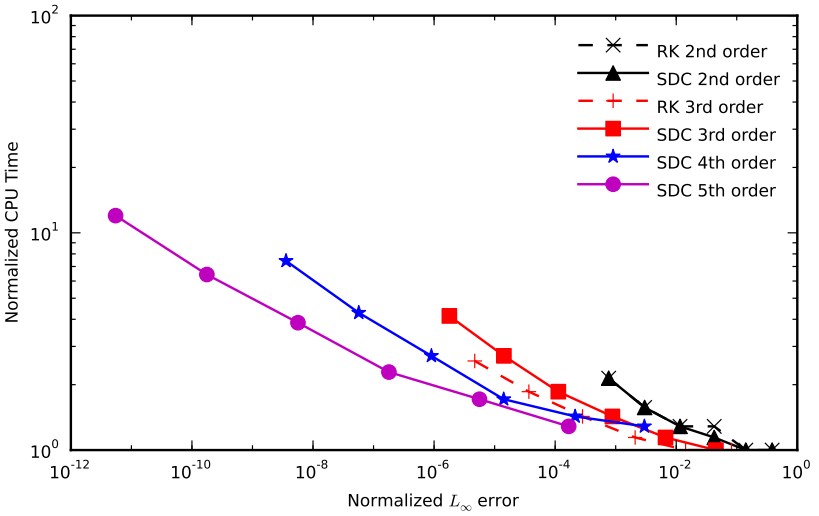


Figure 5. CPU time for different time integration schemes and orders.

of TVD-RK of 2nd and 3rd order, however, RK schemes of 4th order or greater are not TVD either. For certain problems where the error in time is important higher orders may be more suitable. Thus, explicit SDC seems to be a possible approach for high-order time integration of DG schemes.

The test cases below tend to have a dominant spatial error, thus very high orders in time are not required.

4.4. ***Advection of mixed pulses.*** The convection equation (40) is used with the initial value given by

$$u(x,0)=\begin{cases}\frac{1}{6}\left(G(x,\beta,z-\delta)+G(x,\beta,z+\delta)+4G(x,\beta,z)\right) & \text{if } -0.8\le x\le -0.6,\\ 1 & \text{if } -0.4\le x\le -0.2,\\ 1-|10(x-0.1)| & \text{if } \quad 0 \quad \le x\le \quad 0.2,\\ \frac{1}{6}\left(F(x,\alpha,a-\delta)+F(x,\alpha,a+\delta)+4F(x,\alpha,z)\right) & \text{if } \quad 0.4\le x\le \quad 0.6,\\ 0 & \text{otherwise,}\end{cases}$$

for

$$G(x,\beta,z)=e^{-\beta(x-z)^2} \quad \text{and} \quad F(x,\alpha,a)=\sqrt{\max(1-\alpha^2(x-a)^2,0)},$$

with $a=0.5$, $z=-0.7$, $\delta=0.005$, $\alpha=10$, and $\beta=\log 2/(36\delta^2)$. The domain is a uniform grid from $x=-1$ to $x=1$ with periodic boundary conditions.

The SDC method of 3rd order is used. The ML is applied on every element or on the ones flagged by the MB-AP-TVD detector. The result at $t=8.0$ is shown in Figure 6 for $p=2, 4$ and for 200 cells.

Figure 7 shows the L_1 error computed at the center of the elements for different number of cells and polynomial orders:

$$e_{L_1}=\sum_{l=1}^{N}\int_{\Omega_l}|U_l(x,t)-u(x,t)|dx \tag{45}$$

where U is the numerical result and u is the exact solution (or its estimation). In the previous test case L_∞ was used, which is an adequate parameter to analyze smooth solution, however, for discontinuous solutions L_1 is more appropriate.

A few observations can be made from this figure. The error is reduced as the number of elements n or the polynomial order p increases. Also, using the detector improves the accuracy. Note that the order of accuracy is approximately 1 because of the presence of discontinuities in the solution. Therefore, increasing the order p when discontinuities are present reduces the error, but not the order of accuracy.

The same case is run using the original AP-TVD detector, and the efficiency of the AP-TVD and MB-AP-TVD detectors are compared in Figure 8. Curves closer to the bottom left corner represent more accurate schemes.

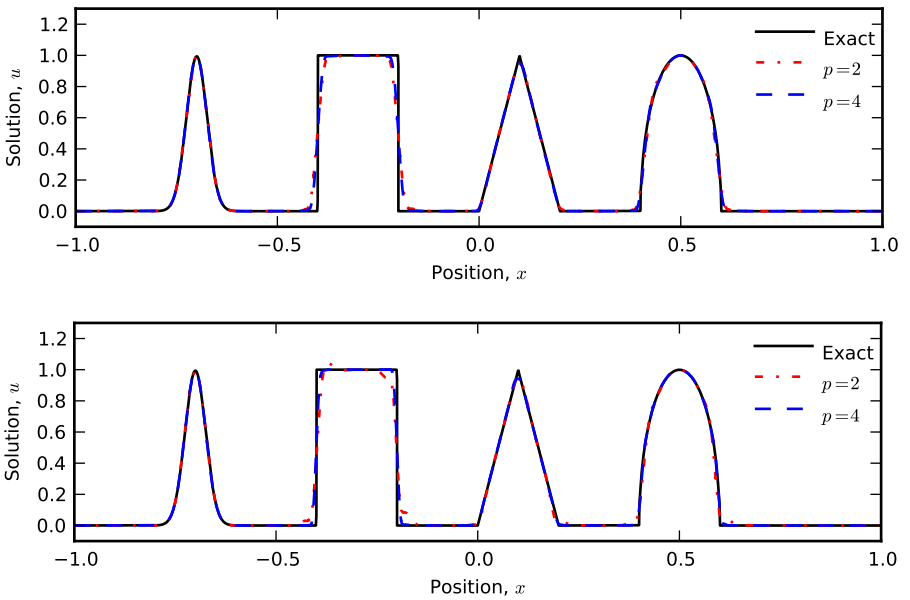


Figure 6. Convection of mixed pulses at $t = 8$, with 200 cells and $p = 2, 4$. Top: ML applied to all cells. Bottom: ML applied only to troubled cells flagged by the detector.

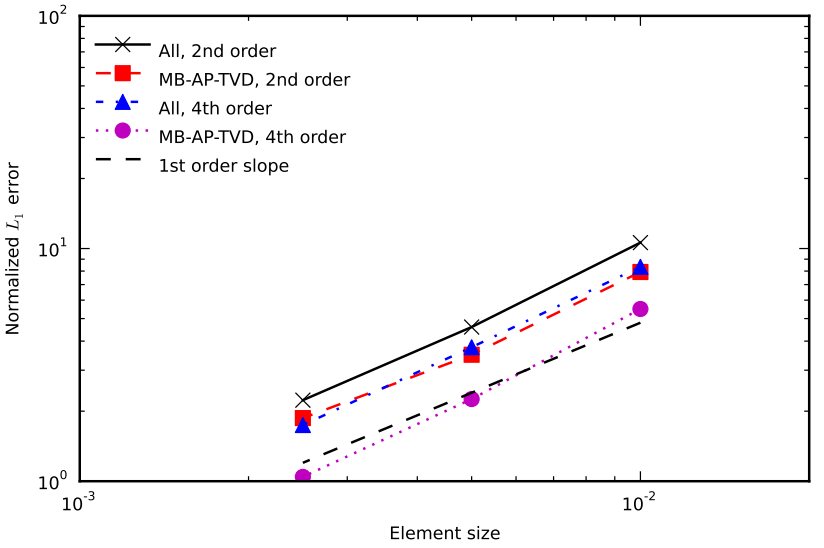


Figure 7. L_1 error for different orders and number of elements for the convection of mixed pulses.

Note that for the same number of elements AP-TVD tends to be slower, while the error is similar. Thus, MB-AP-TVD ends up being a better choice than AP-TVD when the default basis of the element is formed by Legendre polynomials. Given

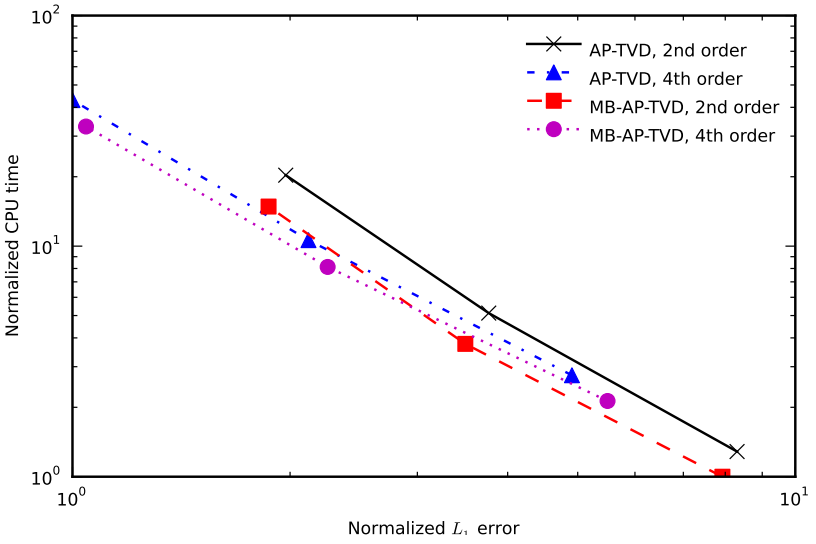


Figure 8. CPU time versus L_1 error for different orders for the convection of mixed pulses.

this, we only use the MB-AP-TVD for the cases below. In addition, Figure 8 shows that for this case, which contains discontinuities, increasing the order of the scheme does not improve its efficiency.

4.5. ***High order for a smooth and nonsmooth solution.*** The same case as in Section 4.2 is observed here for a longer period of time. At $t = 0.4$ a discontinuity is found at approximately $x = 0.1$, while the rest of the solution is smooth. The error is usually computed taking into account the whole domain. However, in order to analyze only the region with a smooth solution, it can be computed for part of the domain. For this purpose we define $\tilde{e}_{L_1}$:

$$\tilde{e}_{L_1} = \sum_{\substack{l=1 \\ 0.3 \le x \le 0.9}}^{N} \int_{\Omega_l} |U_l(x,t) - u(x,t)| dx \tag{46}$$

This is similar to (45), but the integration is done away from the discontinuity, i.e., for $0.3 \le x \le 0.9$. We estimate the exact solution with 512 elements with $p = 6$, and using the SDC of 7th order. The problem is studied using $p = 2, 4, 6$, $N = 10, 20, 30, 40, 80, 160$, SDC of order $p + 1$, and limiting on all the elements or as flagged by the MB-AP-TVD detector. The errors e_{L_1} and $\tilde{e}_{L_1}$ are shown in Tables 4 and 5. Note that the order of accuracy based on $\tilde{e}_{L_1}$ is close to $p+1$, while for e_{L_1} it is close 1. At very low errors the order drops due to accumulated round-off error. For $p = 4$ and $p = 6$ using 20 elements the order is much greater then $p + 1$ since the error for $N = 10$ is relatively large. This is due to the propagation

Number of time steps	Whole domain ($p=2$) e_{L_1}	order	Whole domain ($p=4$) e_{L_1}	order	Whole domain ($p=6$) e_{L_1}	order
10	$3.0333\cdot10^{-2}$	—	$2.6758\cdot10^{-2}$	—	$2.5755\cdot10^{-2}$	—
20	$1.2243\cdot10^{-2}$	1.3089	$1.1347\cdot10^{-2}$	1.2376	$1.1102\cdot10^{-2}$	1.2140
30	$7.6765\cdot10^{-3}$	1.1513	$7.2367\cdot10^{-3}$	1.1093	$7.1097\cdot10^{-3}$	1.0991
40	$5.6299\cdot10^{-3}$	1.0778	$5.3411\cdot10^{-3}$	1.0558	$5.2764\cdot10^{-3}$	1.0366
80	$2.7534\cdot10^{-3}$	1.0319	$2.6630\cdot10^{-3}$	1.0041	$2.0933\cdot10^{-3}$	1.3338
160	$1.3740\cdot10^{-3}$	1.0029	$9.1757\cdot10^{-4}$	1.5372	$1.1018\cdot10^{-3}$	0.9260
Number of time steps	Smooth region ($p=2$) $\tilde{e}_{L_1}$	order	Smooth region ($p=4$) $\tilde{e}_{L_1}$	order	Smooth region ($p=6$) $\tilde{e}_{L_1}$	order
10	$4.4679\cdot10^{-4}$	—	$5.0581\cdot10^{-5}$	—	$3.8820\cdot10^{-5}$	—
20	$2.8444\cdot10^{-5}$	3.9734	$1.3111\cdot10^{-8}$	11.914	$2.1739\cdot10^{-9}$	14.1242
30	$6.2605\cdot10^{-6}$	3.7332	$1.2282\cdot10^{-9}$	5.8399	$2.3454\cdot10^{-12}$	16.8493
40	$2.2165\cdot10^{-6}$	3.6093	$2.3009\cdot10^{-10}$	5.8220	$1.8132\cdot10^{-13}$	8.8985
80	$1.9190\cdot10^{-7}$	3.5298	$5.1326\cdot10^{-12}$	5.4864	$9.6648\cdot10^{-14}$	0.9077
160	$1.7365\cdot10^{-8}$	3.4661	$1.7473\cdot10^{-13}$	4.8765	$1.0902\cdot10^{-13}$	−0.1738

Table 4. The error e_{L_1} (top half) and $\tilde{e}_{L_1}$ (bottom half) with limiting on all the elements.

Number of time steps	Whole domain ($p=2$) e_{L_1}	order	Whole domain ($p=4$) e_{L_1}	order	Whole domain ($p=6$) e_{L_1}	order
10	$2.7784\cdot10^{-2}$	—	$2.4803\cdot10^{-2}$	—	$2.3243\cdot10^{-2}$	—
20	$9.7931\cdot10^{-3}$	1.5044	$7.6011\cdot10^{-3}$	1.7063	$8.0117\cdot10^{-3}$	1.5366
30	$4.4951\cdot10^{-3}$	1.9205	$4.7370\cdot10^{-3}$	1.1663	$4.6348\cdot10^{-3}$	1.3498
40	$4.5546\cdot10^{-3}$	0.0457	$3.2698\cdot10^{-3}$	1.2885	$3.3924\cdot10^{-3}$	1.0848
80	$1.5292\cdot10^{-3}$	1.5745	$1.6939\cdot10^{-3}$	0.9489	$1.7780\cdot10^{-3}$	0.9321
160	$6.5156\cdot10^{-4}$	1.2308	$9.6828\cdot10^{-4}$	0.8068	$9.5344\cdot10^{-4}$	0.8990
Number of time steps	Smooth region ($p=2$) $\tilde{e}_{L_1}$	order	Smooth region ($p=4$) $\tilde{e}_{L_1}$	order	Smooth region ($p=6$) $\tilde{e}_{L_1}$	order
10	$1.5107\cdot10^{-4}$	—	$4.2981\cdot10^{-5}$	—	$1.5876\cdot10^{-5}$	—
20	$9.2714\cdot10^{-6}$	4.0263	$6.3106\cdot10^{-9}$	12.7336	$9.5822\cdot10^{-12}$	20.6600
30	$2.2911\cdot10^{-6}$	3.4477	$6.1355\cdot10^{-10}$	5.7483	$4.4873\cdot10^{-13}$	7.5500
40	$8.4867\cdot10^{-7}$	3.4521	$1.2682\cdot10^{-10}$	5.4799	$1.2121\cdot10^{-13}$	4.5497
80	$8.4769\cdot10^{-8}$	3.3236	$3.3972\cdot10^{-12}$	5.2223	$9.6667\cdot10^{-14}$	0.3265
160	$9.0416\cdot10^{-9}$	3.2289	$1.4398\cdot10^{-13}$	4.5604	$1.0951\cdot10^{-13}$	−0.1799

Table 5. The error e_{L_1} (top half) and $\tilde{e}_{L_1}$ (bottom half) with limiting based on the MB-AP-TVD detector.

to smooth regions of instabilities generated at the discontinuity. For a large enough number of elements, $N \geq 20$, the instabilities do not affect the smooth area being considered in (46) for $\tilde{e}_{L_1}$.

Figure 9 shows the efficiency of the scheme for different orders, with and without the MB-AP-TVD detector. The error and CPU time are normalized based on the case with the largest error.

As observed for previous cases, the troubled-cell detector helps improve the accuracy and efficiency of the solver. For lower L_1 error high-order schemes become more efficient. The limiter reduces numerical oscillations at discontinuities, but with a minimal numerical diffusion, so small instabilities still exist. As the number of elements is increased the numerical error originated at the discontinuity is localized in a smaller region. Thus, probably, p-adaptivity could improve the efficiency by dropping the order at the discontinuity and keeping high order in the smooth region.

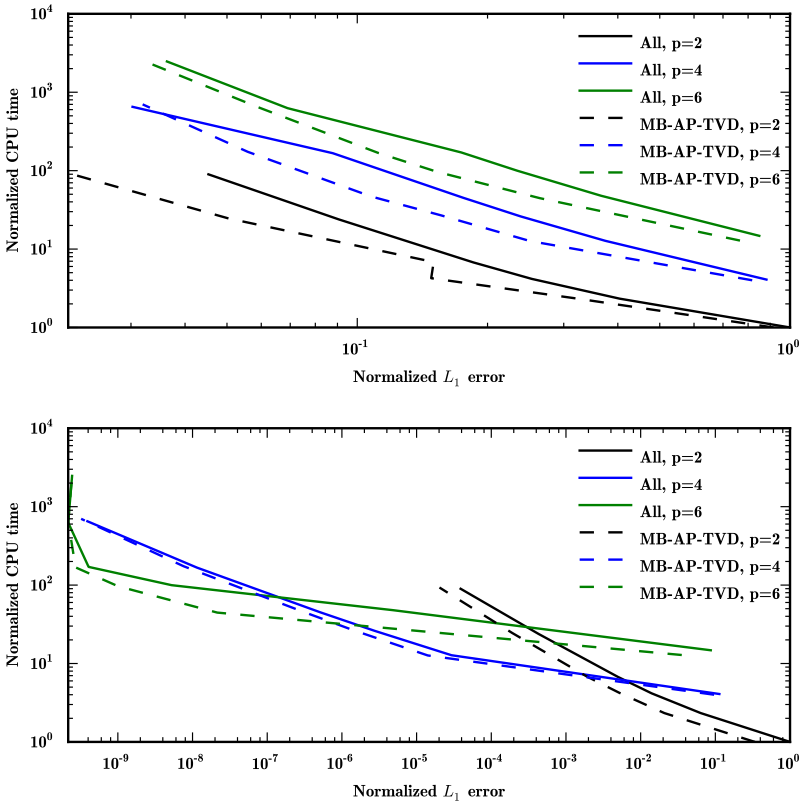


Figure 9. Efficiency of the scheme for different orders for a solution with one discontinuity. The limiter is applied to all the elements or based on the MB-AP-TVD detector. Top: e_{L_1} for the whole domain. Bottom: $\tilde{e}_{L_1}$ for the smooth region.

4.6. ***Sod's problem.*** The initial conditions are

$$(\rho, v, p) = \begin{cases} (1.0, 0.0, 1.0) & \text{if } x \leq 0.5, \\ (0.125, 0.0, 0.1) & \text{if } x > 0.5. \end{cases} \tag{47}$$

A 1D domain is used and it extends from $x = 0$ to $x = 1$. The case is run with different number of elements and the limiting is based on conservative, primitive, or characteristic variables. In addition, two options are tested, one applies the ML with the MB-AP-TVD detector, and the second option applies the ML to all cells. The grid is uniform with $p = 2$, and the time integration is performed using the 3rd-order SDC. The simulation is run until $t = 0.2$. The normalized CPU time versus the L_1 error of the final density is shown for the three cases in Figure 10(a).

Limiting with primitive or characteristic variables requires computing the respective Jacobians for each element, so it is computationally slightly more expensive than using conservative variables, but the error is lower. For primitive and characteristic variables, using the MB-AP-TVD detector to apply the ML to only troubled cells increases the efficiency and lowers the error since the solution is smooth in a large portion of the domain.

The same case is run enabling the adaptive mesh refinement for $\ell_{\max} = 1, 2, 3$. The CPU time versus the L_1 error of the density is shown in Figure 10(b) for limiting with primitive variables. Note that both variables are normalized by the fastest simulation. As $\ell_{\max}$ is increased the curves get slightly closer to the origin. This means that for this test case enabling the adaptivity produces some increase in the efficiency of the solver. Here, the MB-AP-TVD also shows to improve the efficiency.

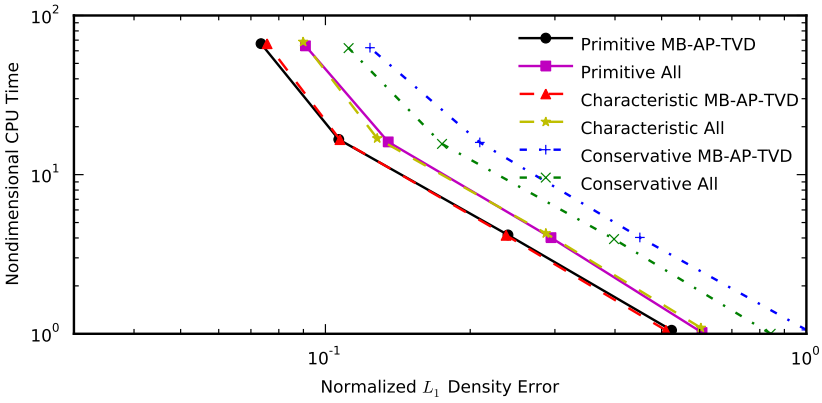
The estimators are compared using $\ell_{\max} = 3$, the MB-AP-TVD detector, and limiting with primitive variables. The efficiency is represented in Figure 10(c). Clearly, JUMP2 is more efficient than KXRCF for this case.

4.7. ***Lax's problem.*** The initial solution is:

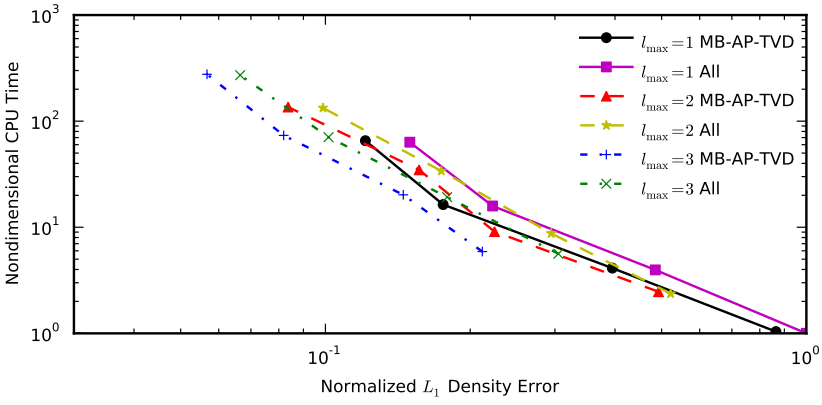
$$(\rho, v, p) = \begin{cases} (0.445, 0.698, 3.528) & \text{if } x \leq 0, \\ (0.5, 0, 0.571) & \text{if } x > 0. \end{cases} \tag{48}$$

The problem is solved in the 1D domain $[-0.5, 0.5]$ until $t = 0.13$.

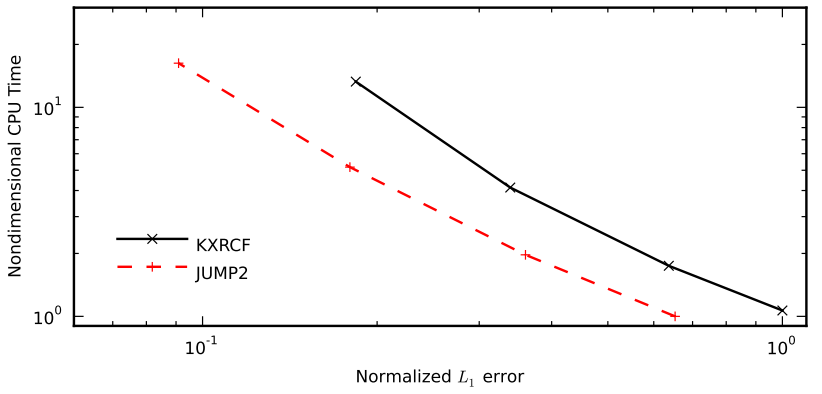
Initially, the effect of the variables used for limiting is studied. A uniform grid is used with $N = 64, 128, 256, 512$ and $p = 2$ with the limiter applied to either all cells or those flagged by the MB-AP-TVD detector. The integration in time is done with the 3rd-order SDC method. The CPU time versus the L_1 error of the density is shown in Figure 11(a). These results show that for this particular test problem the MB-AP-TVD detector increases the efficiency for conservative and characteristic variables, while for primitive variables it did not affect significantly. The CPU time is very similar independent of the set of variables used. Even though



(a) Comparison for different limiting variables with and without detector.

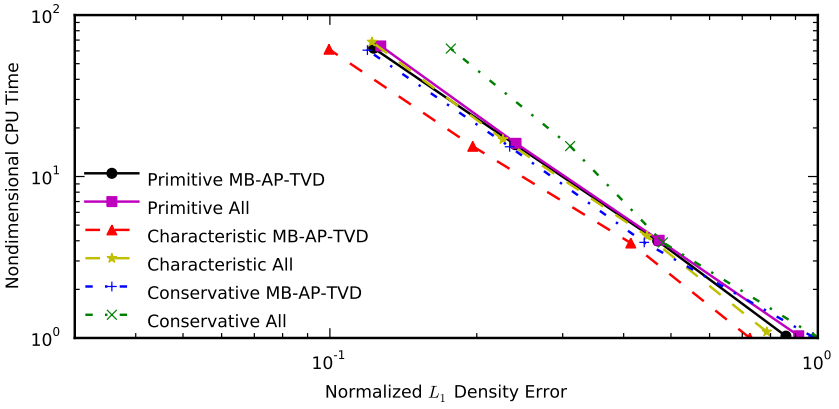


(b) AMR comparison; limiting using primitive variables.

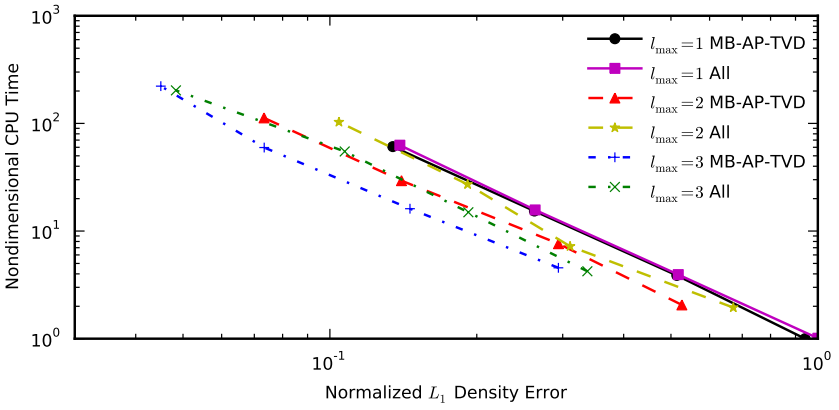


(c) Estimator comparison, limiting using primitive variables; the MB-AP-TVD detector and $\ell_{\max}=3$.

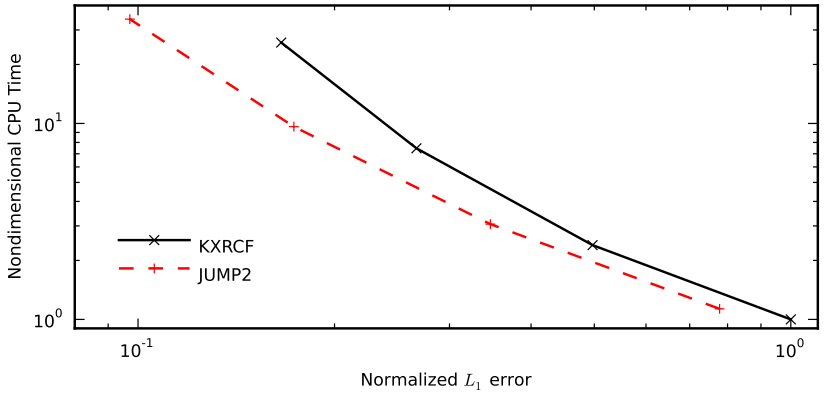
Figure 10. Sod's problem for different solver options. The curves closer to the bottom left corner represent a more efficient set of options.



(a) Limiting comparison.



(b) AMR comparison; limiting using primitive variables.



(c) Estimator comparison; limiting using primitive variables, the MB-AP-TVD detector and $\ell_{\max} = 3$.

Figure 11. Lax problem for different solver options.

conservative variables do not require to compute the Jacobian to transform between the variables, they may require more steps of the limiter.

Now the effect of the adaptation is studied, limiting with primitive variables. The same grid is used, but the adaptation is enable with $\ell_{\max} = 1, 2, 3$. The result is shown in Figure 11(b). When $\ell_{\max}$ is raised, the efficiency of the solver increases and it improves more using the MB-AP-TVD detector. The estimators are compared using $\ell_{\max} = 3$, the MB-AP-TVD detector, and limiting with primitive variables. The efficiency is represented in Figure 11(c). Clearly, JUMP2 is more efficient than KXRCF for this case.

4.8. ***Blast waves.*** Consider the initial data $\rho = 1.0$, $v = 0.0$, and

$$P = \begin{cases} 1000 & \text{if } 0 \le x < 0.1, \\ 0.01 & \text{if } 0.1 \le x < 0.9, \\ 100 & \text{if } 0.9 \le x \le 1.0. \end{cases} \tag{49}$$

This problem is a common test case first presented in [37]. Walls are located at $x = 0$ and $x = 1$.

The problem is run until $t = 0.038$ s for $p = 2$, $\ell_{\max} = 1, 2, 3$, different number of root cells and the 3rd-order SDC. This test case does not have an exact solution, so it is approximated using a uniform mesh with $N = 4096$, $p = 2$, $\ell_{\max} = 1$, the ML without detector and with characteristic decomposition, similar to [21].

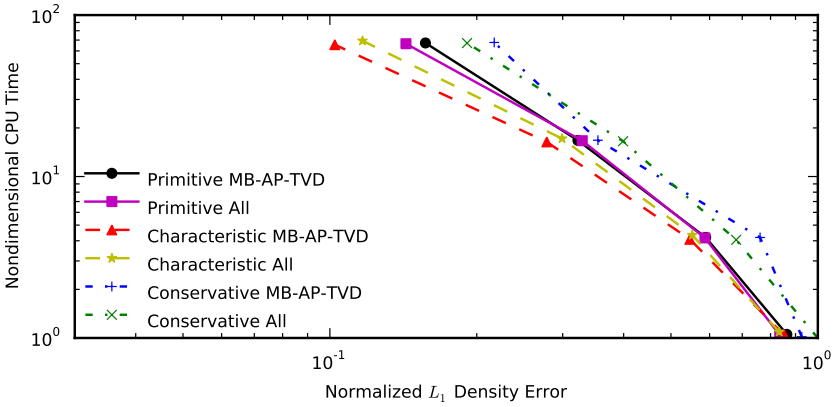
Figure 12 shows the CPU time versus the L_1 error in density, both variables are normalized by the fastest run. Part (a) shows the effect of the detector and the limiting variables. Part (b) represents the efficiency of the AMR approach using primitive variables. The efficiency of the solver clearly improves increasing $\ell_{\max}$. In this case the MB-AP-TVD detector does not produce any significant difference when studying the refinement aspects. The estimators are compared using $\ell_{\max} = 3$, the MB-AP-TVD detector, and limiting with primitive variables and the results are in Figure 12(c). JUMP2 tends to be more efficient than KXRCF for this case.

4.9. ***Shock-entropy wave interaction.*** Consider the Euler equation with the following initial values:

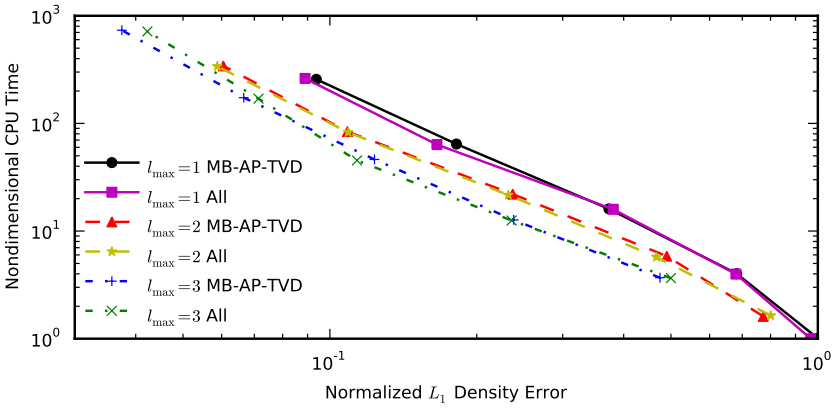
$$(\rho, v, p) = \begin{cases} (3.857143, 2.629369, 10.333333) & \text{if } x < -4, \\ (1.0 + 0.2\sin(5x), 0.0, 1.0) & \text{if } x \ge -4. \end{cases} \tag{50}$$

The problem is solved in the 1D domain $[-5, 5]$ until $t = 1.8$.

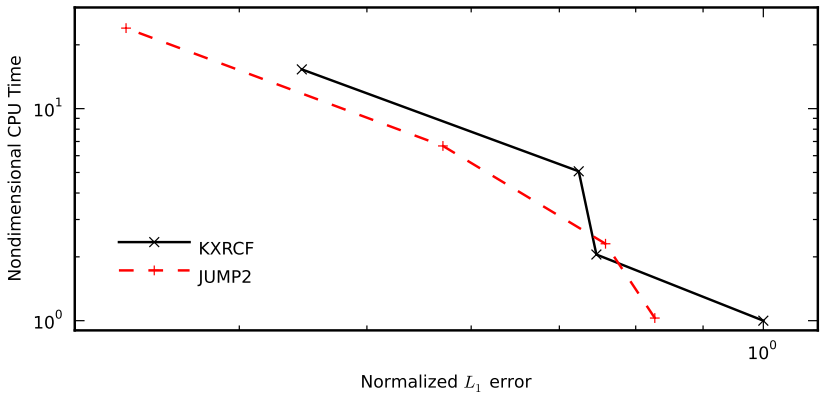
As before, the effect of the variables used for limiting are studied on a uniform grid with $N = 64, 128, 256, 512$ and $p = 2$. The integration in time is done with the 3rd-order SDC method. The CPU time versus the L_1 error of the density is shown in Figure 13. For this problem limiting using primitive variable is advantageous



(a) Limiting comparison.

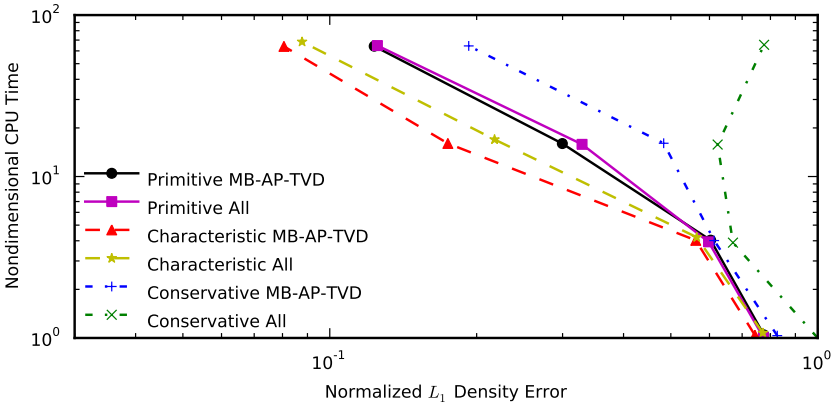


(b) AMR comparison; limiting using primitive variables.

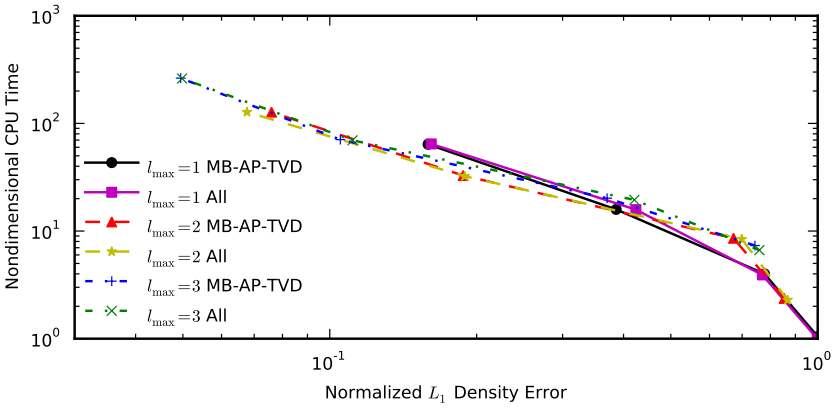


(c) Estimator comparison; limiting using primitive variables, the MB-AP-TVD detector and $\ell_{\max} = 3$.

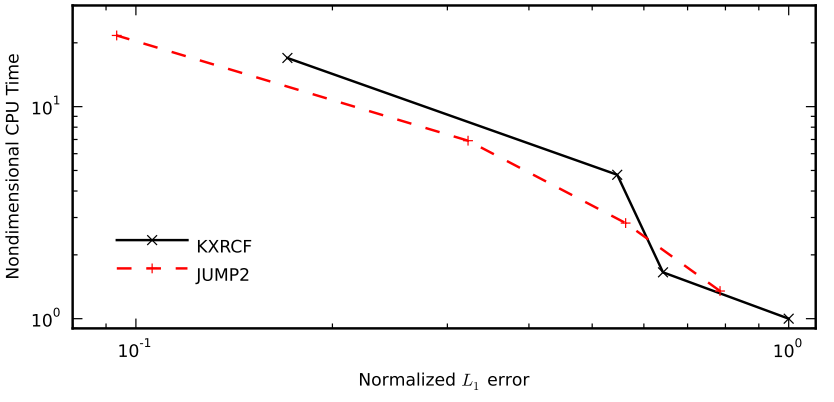
Figure 12. Interacting blast waves for different solver options.



(a) Limiting comparison.



(b) AMR comparison; limiting using the primitive variables.



(c) Estimator comparison; limiting using primitive variables, the MB-AP-TVD detector and $\ell_{\max} = 3$.

Figure 13. Shock-entropy wave interaction problem for different solver options.

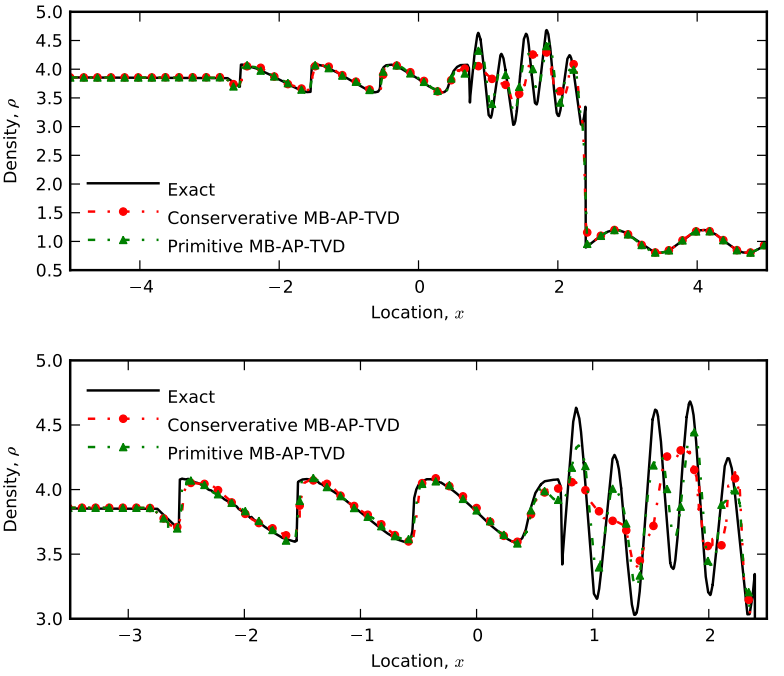


Figure 14. Shock-entropy wave interaction at $t = 1.8$ for $N = 256$, $p = 2$: density ρ as a function of location. The bottom pane shows a detail, to the left of the drop.

compared with conservative variables. Figure 14 shows the solution at $t = 1.8$; it clearly presents that limiting using primitive variables captures the smooth oscillations much more accurately than with conservative variables. Characteristic limiting provides a even more efficient solution than with primitive variables. Also, the MB-AP-TVD detector improves the efficiency. Using AMR for this test case gives no efficiency gains in the low element count (larger normalized error) regime, but AMR is more justified at lower errors where the number of elements increases.

The estimators are compared in Figure 13 using $\ell_{\max} = 3$, the MB-AP-TVD detector, and limiting with primitive variables. JUMP2 tends to be more efficient than KXRCF for this case.

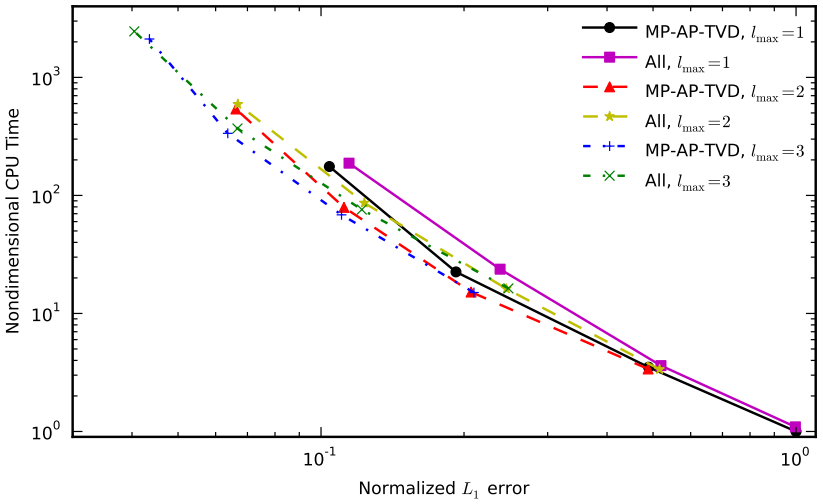
4.10. ***Convection in 2D.*** The limiter and adaptivity approach is studied in 2D using the two-dimensional convection equation

$$\frac{\partial u(\boldsymbol{x}, t)}{\partial t} + c_1 \frac{\partial u(\boldsymbol{x}, t)}{\partial x} + c_2 \frac{\partial u(\boldsymbol{x}, t)}{\partial y} = 0 \tag{51}$$

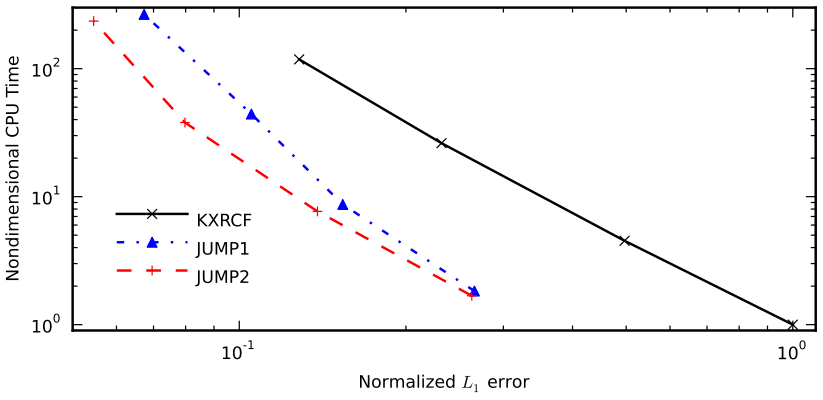
with initial condition

$$u(\boldsymbol{x}, t = 0) = \begin{cases} 1 & \text{for } (x_1 - 0.5)^2 + (x_2 - 0.5)^2 \leq 0.25^2, \\ 0 & \text{otherwise,} \end{cases}$$

where u is a passive scalar, and c_1 and c_2 are the constant convection velocities equal to 1. The domain is the unit square $[0, 1] \times [0, 1]$ with periodic boundary conditions. The time integration used is the 3rd-order SDC. Figure 15 shows the CPU time versus the L_1 error at $t = 1$ for different solver options varying the number of element. In part (a), $\ell_{\max}$ is varied together with the detector. Increasing the $\ell_{\max}$ improves the efficiency, and using the MB-AP-TVD helps too. In part (b), the error estimator for AMR is varied. For this 2D case JUMP1 and JUMP2 produce slightly different results, and JUMP2 is the most efficient of the three estimators.



(a) AMR comparison, with and without detector.



(b) Estimator comparison, using the MB-AP-TVD detector and $\ell_{\max} = 3$.

Figure 15. Two-dimensional convection for different solver options.

4.11. ***Double Mach reflection.*** This is a very common test case for the Euler equation first used by Woodward and Colella [37]. It was also solved by Krivodonova [22] using the ML with a uniform grid and without trouble-cell detector. This case consists of a strong shock impacting a wedge with a half-angle of 30°, thus it is usually simulated by a rectangular domain with a frame rotated 30° over the original horizontal axis.

The rectangular domain has a size of $[0, 4] \times [0, 1]$. A right-moving Mach 10 shock is initially located forming an angle of 60° with the x-axis passing by the coordinate $x = \frac{1}{6}$, $y = 0$. The undisturbed air on the right of the shock has a density of 1.4 and a pressure of 1. The specific heat ratio is $\gamma = 1.4$. A slip-wall boundary is located at the lower boundary from $x = \frac{1}{6}$ to $x = 4$. The right boundary is a supersonic outflow. The left boundary and bottom boundary for $x < \frac{1}{6}$ are supersonic inflow. The reason for applying supersonic inflow at the bottom boundary is to mimic the effect of the wedge. The top boundary mimics the exact motion of the moving shock.

The grid has 48×12 cells with $\ell_{\max} = 5$. The ML is used with the MB-AP-TVD detector. Second-order polynomial elements are used with the 3rd-order SDC method.

The results are shown for $t = 0.2$. Figure 16 shows 60 equally spaced density contours; the inset shows in black the cells flagged by the MB-AP-TVD detector as troubled cells. Note that the ML is not applied here where the flow is uniform, as intended. Figure 17 shows the level of refinement l. It can be noted that the

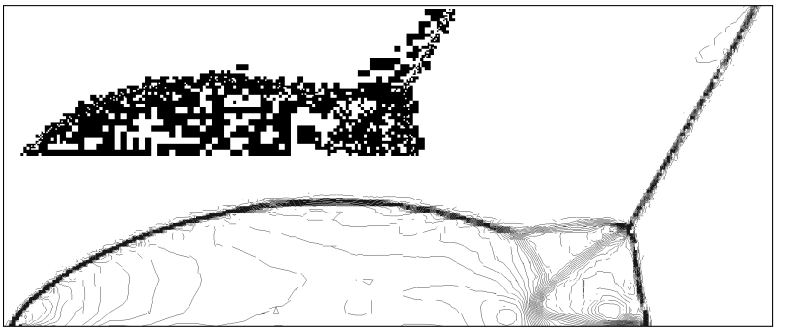


Figure 16. Double Mach reflection: density map at $t = 0.2$ and (inset) troubled cells.

Figure 17. Double Mach reflection: refinement level.

level is increased where the features of the flow are smaller, as can be expected from Figure 16.

4.12. ***Vortex convection.*** As we have seen with previous tests, the global efficiency does not improve significantly for problems dominated by discontinuities when the order is increased. We studied simple smooth cases in 1D, but here we extend the study to a slightly more applicable case in 2D. An isentropic vortex is centered at the center of the domain $(x_c, y_c) = (0.5, 0.5)$. The flow is described by

$$v_1 = M\sqrt{\gamma} + \epsilon\ \tau\ e^{\alpha(1-\tau^2)} \sin\theta, \qquad v_2 = -\epsilon\ \tau\ e^{\alpha(1-\tau^2)} \cos\theta,$$
$$\rho = \left(1 - \frac{\gamma-1}{4\alpha\gamma}\epsilon^2 e^{2\alpha(1-\tau^2)}\right)^{\frac{1}{\gamma-1}}, \qquad p = \rho^\gamma,$$

where $M = 0.3$ and

$$\tau = \frac{r}{r_c}, \qquad r = \sqrt{(x-x_c)^2 + (y-y_c)^2}, \qquad \theta = \arctan\frac{y-y_c}{x-x_c}.$$

Three parameters describe the vortex: its the strength ϵ, its the decay rate α, and the critical radius r_c. For this test the following values are used: $\epsilon = 0.3$, $\alpha = 0.204$, and $r_c = 0.05$. The domain is a unit square with periodic boundaries in every direction. Different number of elements and spatial orders, p, are used. Even though this is a smooth problem, the ML limiter with the MB-AP-TVD detector are used. The range of length scales is very narrow, so AMR is not needed.

Figure 18(a) presents the CPU time versus the L_1 error after one period using the SDC of order $p+1$. The same pattern as for previous 1D cases is observed here. The efficiency increases with the order at in the high accuracy range since at equal CPU time the numerical error is smaller. However, in the low accuracy

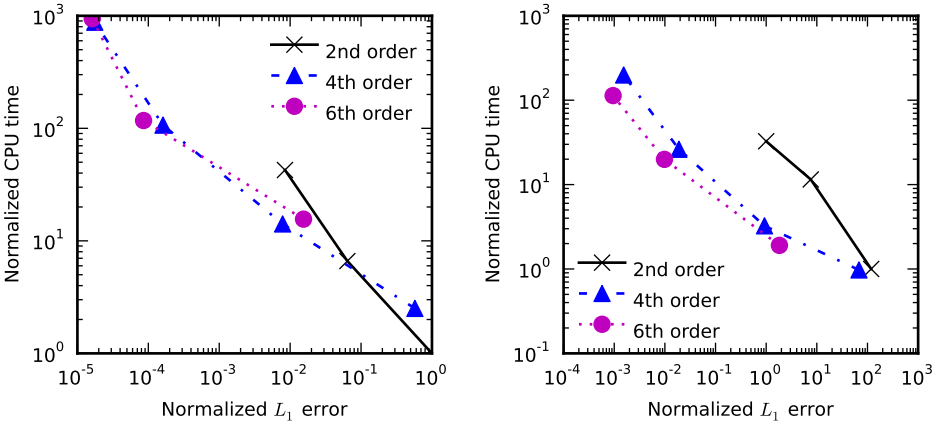


Figure 18. Efficiency for the convection of an isentropic vortex: SDC of order $p+1$ (left) and of order 3 (right).

range low order schemes perform more efficiently. Even though this problem is dominated by convection, the time-step size is limited by the acoustic time, so one could assume that we are over-resolving in time. Thus, we rerun the same cases with the 3rd-order SDC for every p. Note that in this case the CFL has to be adjusted for $p > 2$. We use $C = 0.5$, $C = 0.45$, and $C = 0.4$ for $p = 2$, $p = 4$, and $p = 6$, respectively. Now higher orders in space have a greater advantage. In cases where the error in time is more significant, increasing the order of the scheme in time would make improvements. In this case, however, higher orders in time only add more computational cost.

It can be concluded that the limiting procedure can be freely applied in the whole domain even where smooth features are present. This aspect is important for large-scale applied problems where several types of features can be present at the same time, so a generic and robust scheme is wanted. The shock-vortex interaction case shown below elaborates more on this.

4.13. ***Shock-vortex interaction.*** This problem consists of a vortex going through a shock and helps to test how the solver behaves when smooth features interact with discontinuities. For more information on this kind of problems see [32]. The initial conditions are the same as in [42; 18]. The size of the domain is $[0, 2] \times [0, 1]$. Reflective boundary conditions are used on top and bottom. The left boundary is a supersonic inflow, while the right boundary is an outflow. A stationary shock is located at $x = 0.5$, its preshock Mach number is $M_s = 1.1$, and the left side state is defined by $\rho = 1$, $u = M_s\sqrt{\gamma}$, $v = 0$ and $p = 1$. The right state can be determined from the left state by using the stationary shock relations. An isentropic vortex is centered at $(x_c, y_c) = (0.25, 0.5)$. Therefore, on the left-hand side of the shock the flow is described by

$$v_1 = M_s\sqrt{\gamma} + \epsilon\ \tau\ e^{\alpha(1-\tau^2)} \sin\theta, \qquad v_2 = -\epsilon\ \tau\ e^{\alpha(1-\tau^2)} \cos\theta,$$

$$\rho = \left(1 - \frac{(\gamma-1)\ \epsilon^2\ e^{2\alpha(1-\tau^2)}}{4\alpha\gamma}\right)^{\frac{1}{\gamma-1}}, \quad p = \rho^\gamma$$

where

$$\tau = \frac{r}{r_c} \quad , r = \sqrt{(x-x_c)^2 + (y-y_c)^2}, \quad \theta = \arctan\frac{y-y_c}{x-x_c}.$$

For this test the values used are $\epsilon = 0.3$, $\alpha = 0.204$, and $r_c = 0.05$.

A uniform grid with 32×16 cells and $\ell_{\max} = 4$ is used with $p = 2$. The time integration is done with the 3rd-order SDC method. The ML is used with the MB-AP-TVD detector. The pressure at $t = 0.8$ are shown in Figure 19 with 60 equally spaced contours. The two parts of Figure 20 indicate how the solver adapt to the solution to avoid instabilities and waste computational resources. The vortex successfully goes through the shock and features with different length scales

are properly be resolved. Similar results were observed in [42; 18] using other numerical schemes.

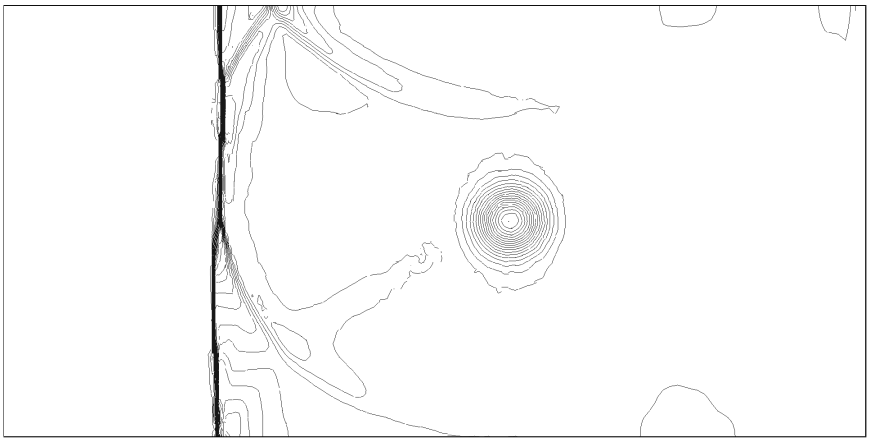


Figure 19. Pressure isocontours for a shock-vortex interaction.

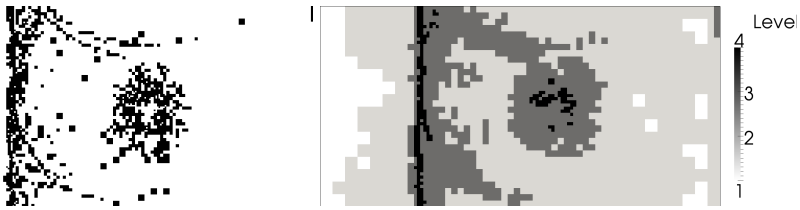


Figure 20. Shock-vortex interaction: troubled cells (left) and refinement level (right)

4.14. ***Spherical shock test.*** The final test is a spherical shock case in a cube defined in $[0, 1] \times [0, 1] \times [0, 1]$. The initial conditions are similar to the typical Sod's problem, but in this case spherical symmetry is used:

$$(\rho, v_1, v_2, v_3, p) = \begin{cases} (1.0, 0.0, 0.0, 0.0, 1.0) & \text{if } r \leq 0.5, \\ (0.125, 0.0, 0.0, 0.0, 0.1) & \text{if } r > 0.5, \end{cases} \tag{52}$$

where r is the distance from $(0, 0, 0)$. The initial grid has 32^3 $p = 2$ elements, each allowed to refine to a level $\ell_{\max} = 3$. The integration in time is done with the 3rd-order SDC method.

An "exact" solution is estimated solving the Euler equation in spherical coordinates assuming spherical symmetry. Thus, the equation being solved in the domain $[0, 1]$ is

$$\frac{\partial \boldsymbol{u}}{\partial t} + \frac{\partial \boldsymbol{F}(\boldsymbol{u})}{\partial x} = \boldsymbol{S}, \tag{53}$$

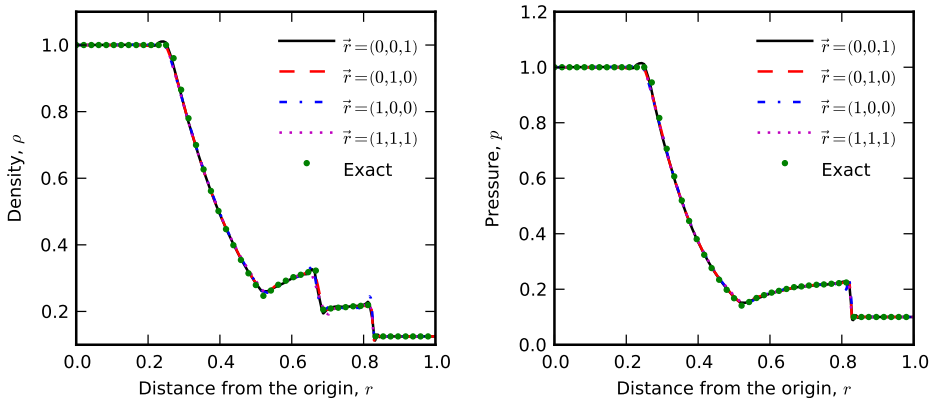


Figure 21. Spherical shock test at $t = 0.15$ over four different vectors.

where $\boldsymbol{S} = -2/x\left(\rho v_1, \rho v_1^2, (\rho E_T + p)v_1\right)^T$. This 1D problem is solved on a grid with 1024 cells with $p = 2$ and integrated in time with the 3rd-order SDC method.

A very similar test case to this one was studied in 2D in [39; 36].

The density and pressure at $t = 0.2$ over four different vectors are shown in Figure 21. Each of these four vectors are: (1, 0, 0), (0, 1, 0), (0, 0, 1), and (1, 1, 1). Given that the density on the different trajectories match, the scheme successfully respects the spherical symmetry of the problem. Note that the results shown do not match exactly the classical one-dimensional Sod shock-tube problem due to 3D effects. Figure 22 on the next page demonstrates the ability of AMR to track the shock and rarefaction waves as required.

5. Conclusions

The Euler equations are solved using the discontinuous Galerkin method with adaptive mesh refinement and high-order of accuracy in space and time.

It was shown using high-order schemes that problems with discontinuities can present high order of convergence in the smooth regions, while the global order of accuracy is close to 1 in the L_1 and L_∞ norm. Most of the cases studied include discontinuities. Therefore, in such cases the order in space and time of the scheme used is 3, since, as it was shown, increasing the order of the solver does not improve its efficiency significantly when discontinuities dominate. Given that the time step is limited by the acoustic time, convection-dominated problems end up being over-resolved in time, so in such cases increasing the order in time produces an unnecessary computational cost.

A simple and effective error estimator for adaptivity based on the interelement jump is suggested and it was shown to be more efficient than other estimator found in the literature. From a computational-resources point of view, the most

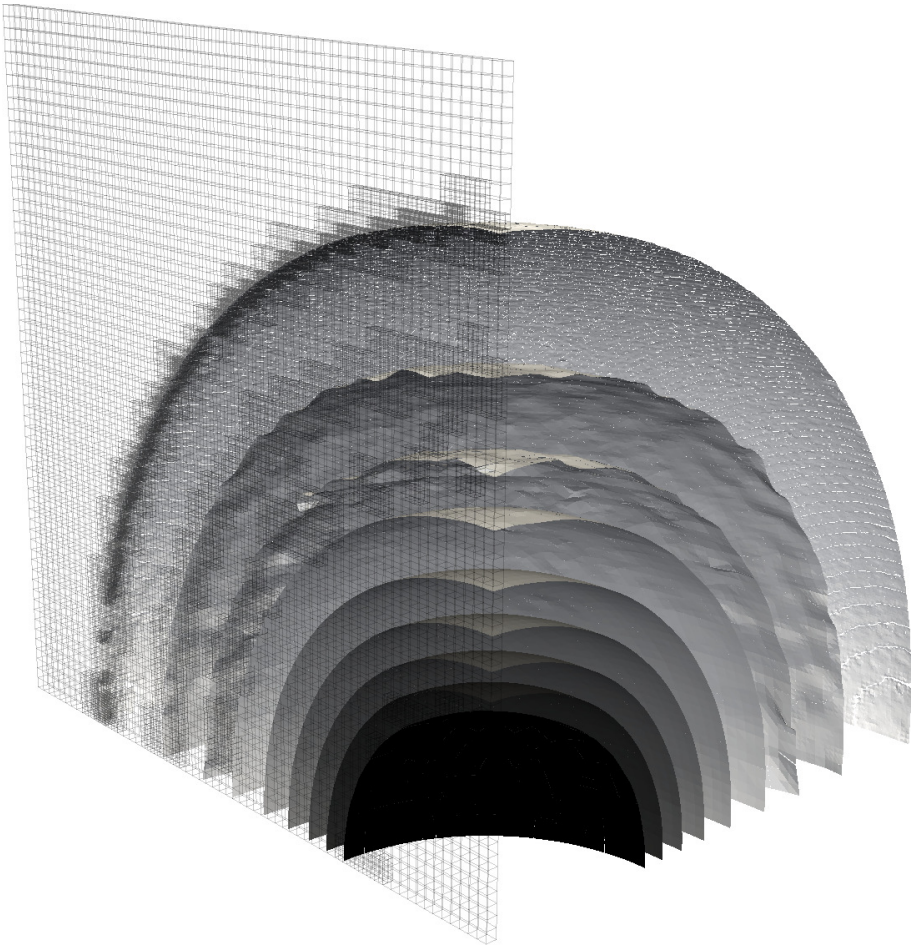


Figure 22. Spherical shock test: density contours and grid refinement.

efficient combination of maximum levels of refinement and initial number of cells is problem-dependent. In a few of the tested problems the overhead caused by the adaptation made it unnecessary. However, in no case with a wide range of scales AMR caused a significant loss of efficiency.

The AP-TVD detector in [42] was modified replacing the averaged-derivative basis that it originally required by the Legendre polynomial basis, which is commonly used in DG. Therefore, the current approach avoids the transformation and a better efficiency of the scheme is observed. We named it the moment-based AP-TVD (MB-AP-TVD) since it uses the default moments of the solution — like the moment limiter (ML) does. Yang and Wang [42] showed that the AP-TVD detector gives better results than the more common detectors used in [29], so the MB-AP-TVD should be even more efficient than those.

The troubled cells were treated with a ML modified for nonuniform meshes with hanging nodes. The limiting stage is done using primitive, characteristic, and conservative variables and then appropriately evaluated. The optimal choice of limiting variables and where to apply the limiter is case-specific, but based on the results of the one-dimensional tests limiting using primitive variables and the MB-AP-TVD detector is the recommended starting point, especially for multidimensional problems since the ML is inherently multidimensional and the characteristic decomposition, slightly better than primitive variables in 1D, cannot be applied in 2 or 3 dimensions. The computational cost due to the conversion from conservative to the other variables seems to be negligible. This Jacobian (and its inverse) is computed each time an element is being limited, but the CPU advantage of conservative variables seems to be lost since worse limiting requires more correction steps of the limiter.

In addition, most test cases were studied with SDC method, what shows that it is an adequate time-integration scheme that could be considered as an alternative to the Runge–Kutta methods for certain applications, especially as the order increases since it is easier to derive and implement. More research is still necessary to determine the numerical properties of SDC-DG, such as its maximum CFL number and its numerical dissipation at different frequencies. For cartesian, low-order cases DG may perform similarly to FD or FV [44]. However, it is important to note that when the conditions are more sophisticated (e.g., unstructured, noncartesian, high-order), where other schemes cannot even be applied, DG still performs well.

The scheme, including our new developments, are relatively simple to implement, robust, with great numerical properties. Thus, it presents a technique that should be exploited for more generic applications.

For steady-state problems the proposed approach may not be highly efficient. Common modifications to improve the convergence to a steady state include some type of filter in time for the limiter and other discrete operations, since they create oscillations that do not let the residual decrease enough. However, the goal of this study is to investigate methods needed for time dependent problems.

p-adaptivity could be useful especially for problems with discontinuities, which are better treated with low-order schemes. Application to the full Navier–Stokes equation will be reported in the near future.

Acknowledgements

The authors wish to thank Prof. Yingjie Liu of School of Mathematics, Georgia Institute of Technology for his helpful suggestions. This work is supported in part by General Electric Infrastructure – Aviation. Also, the authors are thankful for observations made by the reviewers.

Appendix: Moment limiter in two and three dimensions

The 1D momentum limiter presented in Section 3.4 is discussed here in two and three dimensions for completeness.

A.1. ***Two-dimensional moment limiter.*** In this case cross derivatives should be taken into account. Hence, for element l, m,

$$\frac{\partial^{i+j}\tilde{U}_{l,m}}{\partial x_1^i \partial x_2^j} = \text{minmod}\left(\frac{\partial^{i+j}U_{l,m}}{\partial x_1^i \partial x_2^j}, \beta_{ij} D_{ij}^{x_1+}, \beta_{ij} D_{ij}^{x_1-}, \beta_{ij} D_{ij}^{x_2+}, \beta_{ij} D_{ij}^{x_2-}\right) \quad (54)$$

where the frame (x_1, x_2) is a rotation of (x, y) aligned to the computational coordinates (ξ, η) of the current element.

In this case, the limiting starts from orders (p, p), and continuous with the pair $(p, p-1)$ and $(p-1, p)$, then with the pair $(p, p-2)$ and $(p-2, p)$, and so on until $(p, 0)$ and $(0, p)$. Then the loop starts again from $(p-1, p-1)$, and continuous with $(p-1, p-2)$ and $(p-2, p-1)$, and so on. Whenever a pair is not changed the limiting procedure is stopped.

If a neighboring cell is split because of refinement, the average between the two neighboring children is used. If a neighboring cell is coarser because the current cell is more refined, the modes of the neighbor have to be computed as it were refined too. Note that the characteristic decomposition is only consistent in a 1D sense. Given that the ML can be multidimensional, the characteristic decomposition would have to be done in an arbitrary direction. Therefore, for multidimensional cases a primitive-variable decomposition may be more appropriate.

A.2. ***Three-dimensional moment limiter.*** In this case, for element l, m, n,

$$\frac{\partial^{i+j+k}\tilde{U}_{l,m,n}}{\partial x_1^i \partial x_2^j \partial x_3^k} = \text{minmod}\left(\frac{\partial^{i+j+k}U_{l,m,n}}{\partial x_1^i \partial x_2^j \partial x_3^j}, \beta_{ijk} D_{ijk}^{x_1+}, \beta_{ijk} D_{ijk}^{x_1-}, \right.$$
$$\left. \beta_{ijk} D_{ijk}^{x_2+}, \beta_{ijk} D_{ijk}^{x_2-}, \beta_{ijk} D_{ijk}^{x_3+}, \beta_{ijk} D_{ijk}^{x_3-}\right), \quad (55)$$

where the frame (x_1, x_2, x_3) is a rotation of (x, y, z) aligned to the computational coordinates (ξ, η, ζ) of the current element.

In this case, the limiting starts from orders (p, p, p), and continuous for the triad $(p, p, p-1)$, $(p, p-1, p)$ and $(p-1, p, p)$, then for the triad $(p, p, p-2)$, $(p, p-2, p)$ and $(p-2, p, p)$, and so on until $(p, p, 0)$, $(p, 0, p)$ and $(0, p, p)$. Then the loop starts again from $(p-1, p-1, p-1)$, and continues for $(p-1, p-1, p-2)$, $(p-1, p-2, p-1)$ and $(p-2, p-1, p-1)$, and so on. Whenever a triad is not changed the limiting procedure is stopped.

If a neighboring cell is split because of refinement, the average between the four neighboring children is used. Like in the 2D case, a primitive-variable decomposition may be the optimal approach.

References

[1] G. E. Barter and D. L. Darmofal, *Shock capturing with higher-order, PDE-based artificial viscosity*, 18th AIAA Computational Fluid Dynamics Conference (2007-3823), AIAA, 2007.

[2] R. Biswas, K. D. Devine, and J. E. Flaherty, *Parallel, adaptive finite element methods for conservation laws*, Appl. Numer. Math. **14** (1994), no. 1-3, 255–283. MR 1273828 Zbl 0826.65084

[3] B. Cockburn, S. Hou, and C.-W. Shu, *The Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws, IV: The multidimensional case*, Math. Comp. **54** (1990), no. 190, 545–581. MR 90k:65162

[4] B. Cockburn, S. Y. Lin, and C.-W. Shu, *TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws, III: One-dimensional systems*, J. Comput. Phys. **84** (1989), no. 1, 90–113. MR 90k:65161

[5] B. Cockburn and C.-W. Shu, *TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws, II: General framework*, Math. Comp. **52** (1989), no. 186, 411–435. MR 90k:65160

[6] ———, *The Runge–Kutta discontinuous Galerkin method for conservation laws, V: Multidimensional systems*, J. Comput. Phys. **141** (1998), no. 2, 199–224. MR 99c:65181 Zbl 0920.65059

[7] ———, *Runge–Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput. **16** (2001), no. 3, 173–261. MR 2002i:65099 Zbl 1065.76135

[8] A. Dutt, L. Greengard, and V. Rokhlin, *Spectral deferred correction methods for ordinary differential equations*, BIT **40** (2000), no. 2, 241–266. MR 2001e:65104 Zbl 0959.65084

[9] J. E. Flaherty, L. Krivodonova, J.-F. Remacle, and M. S. Shephard, *Aspects of discontinuous Galerkin methods for hyperbolic conservation laws*, Finite Elem. Anal. Des. **38** (2002), no. 10, 889–908. MR 2003e:65176 Zbl 0996.65106

[10] F. X. Giraldo, J. S. Hesthaven, and T. Warburton, *Nodal high-order discontinuous Galerkin methods for the spherical shallow water equations*, J. Comput. Phys. **181** (2002), no. 2, 499–525. MR 2003g:86004 Zbl 1178.76268

[11] S. Gottlieb, D. I. Ketcheson, and C.-W. Shu, *High order strong stability preserving time discretizations*, J. Sci. Comput. **38** (2009), no. 3, 251–289. MR 2010b:65161 Zbl 1203.65135

[12] S. Gottlieb and C.-W. Shu, *Total variation diminishing Runge–Kutta schemes*, Math. Comp. **67** (1998), no. 221, 73–85. MR 98c:65122 Zbl 0897.65058

[13] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev. **43** (2001), no. 1, 89–112. MR 2002f:65132 Zbl 0967.65098

[14] J. Grooss and J. S. Hesthaven, *A level set discontinuous Galerkin method for free surface flows*, Comput. Methods Appl. Mech. Engrg. **195** (2006), no. 25-28, 3406–3429. MR 2007e:76204 Zbl 1121.76035

[15] A. Harten, *ENO schemes with subcell resolution*, J. Comput. Phys. **83** (1989), no. 1, 148–184. MR 90i:76010 Zbl 0696.65078

[16] J. Jaffré, C. Johnson, and A. Szepessy, *Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws*, Math. Models Methods Appl. Sci. **5** (1995), no. 3, 367–386. MR 96c:65164 Zbl 0834.65089

[17] B. Jeon, J. D. Kress, L. A. Collins, and N. Grønbech-Jensen, *Parallel tree code for two-component ultracold plasma analysis*, Comput. Phys. Commun. **178** (2008), 272–279.

[18] G.-S. Jiang and C.-W. Shu, *Efficient implementation of weighted ENO schemes*, J. Comput. Phys. **126** (1996), no. 1, 202–228. MR 97e:65081 Zbl 0877.65065

[19] A. M. Khokhlov, *Fully threaded tree algorithms for adaptive refinement fluid dynamics simulations*, J. Comput. Phys. **143** (1998), no. 2, 519–543. MR 1631200

[20] R. M. Kirby and G. E. Karniadakis, *De-aliasing on non-uniform grids: algorithms and applications*, J. Comput. Phys. **191** (2003), 249–264. Zbl 1161.76534

[21] L. Krivodonova, J. Xin, J.-F. Remacle, N. Chevaugeon, and J. E. Flaherty, *Shock detection and limiting with discontinuous Galerkin methods for hyperbolic conservation laws*, Appl. Numer. Math. **48** (2004), no. 3-4, 323–338. MR 2056921 Zbl 1038.65096

[22] L. Krivodonova, *Limiters for high-order discontinuous Galerkin methods*, J. Comput. Phys. **226** (2007), no. 1, 879–896. MR 2008j:65162 Zbl 1125.65091

[23] T. Leicht and R. Hartmann, *Error estimation and anisotropic mesh refinement for 3d laminar aerodynamic flow simulations*, J. Comput. Phys. **229** (2010), no. 19, 7344–7360. MR 2011k:76054 Zbl 05785978

[24] Y. Liu, C.-W. Shu, E. Tadmor, and M. Zhang, *Central discontinuous Galerkin methods on overlapping cells with a nonoscillatory hierarchical reconstruction*, SIAM J. Numer. Anal. **45** (2007), no. 6, 2442–2467. MR 2009a:65256 Zbl 1157.65450

[25] Y. Liu, C.-W. Shu, and Z. Xu, *Hierarchical reconstruction with up to second degree remainder for solving nonlinear conservation laws*, Nonlinearity **22** (2009), no. 12, 2799–2812. MR 2011c:35338 Zbl 1184.65086

[26] Y. Liu, C.-W. Shu, and M. Zhang, *Strong stability preserving property of the deferred correction time discretization*, J. Comput. Math. **26** (2008), no. 5, 633–656. MR 2010f:65161 Zbl 1174.65036

[27] M. L. Minion, *Semi-implicit spectral deferred correction methods for ordinary differential equations*, Commun. Math. Sci. **1** (2003), no. 3, 471–500. MR 2005f:65085 Zbl 1088.65556

[28] P.-O. Persson and J. Peraire, *Sub-cell shock capturing for discontinuous Galerkin methods*, 44th AIAA Aerospace Sciences Meeting and Exhibit, AIAA, January 2006.

[29] J. Qiu and C.-W. Shu, *A comparison of troubled-cell indicators for Runge–Kutta discontinuous Galerkin methods using weighted essentially nonoscillatory limiters*, SIAM J. Sci. Comput. **27** (2005), no. 3, 995–1013. MR 2006j:65293 Zbl 1092.65084

[30] ———, *Hermite WENO schemes and their application as limiters for Runge–Kutta discontinuous Galerkin method, II: Two dimensional case*, Comput. & Fluids **34** (2005), no. 6, 642–663. MR 2005j:65086

[31] ———, *Runge–Kutta discontinuous Galerkin method using WENO limiters*, SIAM J. Sci. Comput. **26** (2005), no. 3, 907–929. MR 2005j:65088 Zbl 1077.65109

[32] A. Rault, G. Chiavassa, and R. Donat, *Shock-vortex interactions at high Mach numbers*, J. Sci. Comput. **19** (2003), no. 1-3, 347–371. MR 2004i:76145 Zbl 1039.76047

[33] J.-F. Remacle, J. E. Flaherty, and M. S. Shephard, *An adaptive discontinuous Galerkin technique with an orthogonal basis applied to compressible flow problems*, SIAM Rev. **45** (2003), no. 1, 53–72. MR 2004k:65176 Zbl 1127.65323

[34] J.-F. Remacle, X. Li, M. S. Shephard, and J. E. Flaherty, *Anisotropic adaptive simulation of transient flows using discontinuous Galerkin methods*, Internat. J. Numer. Methods Engrg. **62** (2005), no. 7, 899–923. MR 2005h:76057 Zbl 1078.76042

[35] C.-W. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock-capturing schemes. II*, J. Comput. Phys. **83** (1989), no. 1, 32–78. MR 90i:65167 Zbl 0674.65061

[36] E. F. Toro, *Riemann solvers and numerical methods for fluid dynamics: A practical introduction*, 2nd ed., Springer, Berlin, 1999. MR 2000f:76091 Zbl 0923.76004

[37] P. Woodward and P. Colella, *The numerical simulation of two-dimensional fluid flow with strong shocks*, J. Comput. Phys. **54** (1984), no. 1, 115–173. MR 85e:76004 Zbl 0573.76057

[38] Y. Xia, Y. Xu, and C.-W. Shu, *Efficient time discretization for local discontinuous Galerkin methods*, Discrete Contin. Dyn. Syst. Ser. B **8** (2007), no. 3, 677–693. MR 2008e:65307 Zbl 1141.65076

[39] J. Xin and J. E. Flaherty, *Viscous stabilization of discontinuous Galerkin solutions of hyperbolic conservation laws*, Appl. Numer. Math. **56** (2006), no. 3-4, 444–458. MR 2006j:65296 Zbl 1089.65101

[40] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for high-order time-dependent partial differential equations*, Commun. Comput. Phys. **7** (2010), no. 1, 1–46. MR 2011g:65204

[41] Z. Xu, Y. Liu, and C.-W. Shu, *Hierarchical reconstruction for discontinuous Galerkin methods on unstructured grids with a WENO-type linear reconstruction and partial neighboring cells*, J. Comput. Phys. **228** (2009), no. 6, 2194–2212. MR 2010b:65213 Zbl 1165.65392

[42] M. Yang and Z. J. Wang, *A parameter-free generalized moment limiter for high-order methods on unstructured grids*, Adv. Appl. Math. Mech. **1** (2009), no. 4, 451–480. MR 2010h:65182

[43] L. Yuan and C.-W. Shu, *Discontinuous Galerkin method based on non-polynomial approximation spaces*, J. Comput. Phys. **218** (2006), no. 1, 295–323. MR 2008c:65267 Zbl 1104.65094

[44] T. Zhou, Y. Li, and C.-W. Shu, *Numerical comparison of WENO finite volume and Runge–Kutta discontinuous Galerkin methods*, J. Sci. Comput. **16** (2001), no. 2, 145–171. MR 2002k:65133 Zbl 0991.65083

[45] H. Zhu and J. Qiu, *Adaptive Runge–Kutta discontinuous Galerkin methods using different indicators: one-dimensional case*, J. Comput. Phys. **228** (2009), no. 18, 6957–6976. MR 2011a:65339 Zbl 1173.65339

Received December 15, 2010. Revised April 8, 2012.

LEANDRO D. GRYNGARTEN: leandro@gatech.edu
Aerospace Engineering, Georgia Institute of Techonology, 270 Ferst Drive, Atlanta, GA 30332, United States

ANDREW SMITH: andrew.g.smith@gatech.edu
Aerospace Engineering, Georgia Institute of Techonology, 270 Ferst Drive, Atlanta, GA 30332, United States

SURESH MENON: suresh.menon@ae.gatech.edu
Aerospace Engineering, Georgia Institute of Techonology, 270 Ferst Drive, Atlanta, GA 30332, United States

COMM. APP. MATH. AND COMP. SCI.
Vol. 7, No. 2, 2012
dx.doi.org/10.2140/camcos.2012.7.175

msp

ANALYSIS OF PERSISTENT NONSTATIONARY TIME SERIES AND APPLICATIONS

PHILIPP METZNER, LARS PUTZIG AND ILLIA HORENKO

We give an alternative and unified derivation of the general framework developed in the last few years for analyzing nonstationary time series. A different approach for handling the resulting variational problem numerically is introduced. We further expand the framework by employing adaptive finite element algorithms and ideas from information theory to solve the problem of finding the most adequate model based on a maximum-entropy ansatz, thereby reducing the number of underlying probabilistic assumptions. In addition, we formulate and prove the result establishing the link between the optimal parametrizations of the direct and the inverse problems and compare the introduced algorithm to standard approaches like Gaussian mixture models, hidden Markov models, artificial neural networks and local kernel methods. Furthermore, based on the introduced general framework, we show how to create new data analysis methods for specific practical applications. We demonstrate the application of the framework to data samples from toy models as well as to real-world problems such as biomolecular dynamics, DNA sequence analysis and financial applications.

1. Introduction

In the field of time series analysis, a common problem is the analysis of high-dimensional time series containing possibly hidden information at different time scales. Here we consider the analysis of *persistent processes*, those where the temporal change of the underlying model parameters takes place at a much slower pace than the change of the system variables themselves. Such systems could be financial markets (where the underlying dynamics might drastically change due to

Illia Horenko is the corresponding author.

Work supported by the Swiss National Science Foundation (project "AnaGraph"), the German DFG SPP 1276 (MetStröm) "Meteorology and Turbulence Mechanics" and by the Swiss HP2C initiative "Swiss Platform for High-Performance and High-Productivity Computing". P. Metzner acknowledges the financial support of the DFG priority programme 1276 (MetStröm) "Multiple Scales in Fluid Mechanics and Meteorology".

MSC2010: primary 60G20, 62H25, 62H30, 62M10, 62M20; secondary 62M07, 62M09, 62M05, 62M02.

Keywords: nonstationary time series analysis, nonstationary data analysis, clustering, finite element method.

market breakdowns, new laws, etc.) [27; 48; 63]; climate systems (depending on the external factors like insolation, human activity, etc.) [54; 18; 39; 38; 32; 30]; ocean circulation models [22; 23] or biophysical systems [67; 37; 36; 41; 62; 68].

In the literature, the problem of data-based phase identification is addressed by a huge number of approaches which can be roughly classified as either *non-dynamical* or *dynamical* methods. The class on nondynamical methods exploits solely *geometrical* properties of the data for clustering regardless of their temporal occurrence. The most prominent approach is the k-means method [53], which clusters data points according to their minimal distance to geometrical centroids of point clouds.

Dynamical methods additionally take into account the temporal dynamics of data. This class of methods can further be divided into Bayesian approaches, such as the hidden Markov model (HMM) [4; 3; 56; 37; 36] or the Gaussian mixture model (GMM) (see, e.g., [21]) and the so-called *local kernel methods* (moving window methods) [20; 52]. Although the Bayesian methods have proven to be very successful in applications ranging from speech recognition [64] over atmospheric flows identification [54; 18] to conformation dynamics of biomolecules [17], they are based on the restrictive assumption that the underlying dynamics are governed by a *stationary probabilistic model*. Particularly, the assumption of stationarity implies, e.g., a locally constant mean value and a locally constant variance. In many real world applications, however, these implications are not valid due to theoretical reasons or simply due to the lack of sufficiently long time series of observations.

In local kernel methods the assumption of stationarity is relaxed by applying *nonparametric regression methods* to estimate time-dependent statistical properties of the underlying data. The key idea is the following: instead of considering every element of the time series to be equally statistically important, for a fixed time t the data is *weighted* with a suitable so-called *kernel* function, e.g., a Gaussian probability density function. The modified time series then is considered to be stationary and, consequently, statistical objects can be computed by standard procedures.

The nonstationary time series analysis methods that have been developed in the group of I. Horenko and that will be considered in the current manuscript can be seen as a generalization of the idea described above. Therefore we explain the procedure in more detail. Suppose we observed a time series of real-valued observations discretely in time, denoted by $\boldsymbol{X} = (x_{t_0}, \ldots, x_{t_T})$ with $0 \leq t_0 < \ldots < t_T \leq 1$. Further suppose that the time series is appropriately described by the model

$$x_{t_i} = \mu(t_i) + \varepsilon_{t_i}, \quad i = 0, \ldots, T, \tag{1}$$

where $\{\varepsilon_{t_i}\}$ is a family of independent and identically distributed (i.i.d.) random variables with $\mathbb{E}[\varepsilon(t_i)] = 0$. An estimator for $\mu(t),\ t \in [0, 1]$ is given by [19; 20]

$$\hat{\mu}(t) = \frac{1}{b}\sum_{j=0}^{T} x_{t_j} \int_{s_j}^{s_{j+1}} W\left(\frac{t-s}{b}\right) \mathrm{d}s, \tag{2}$$

where $W(\cdot)$ is a nonnegative kernel function satisfying the conditions

$$\int_{-\infty}^{\infty} W(s)\,\mathrm{d}s = 1, \quad \int_{-\infty}^{\infty} (W(s))^2\,\mathrm{d}s < \infty \tag{3}$$

and $0 = s_0 \le t_0 \le s_1 \le t_1 \le \ldots \le t_T \le s_{T+1} = 1$. The parameter $b \in \mathbb{R}$ is referred to as the window size associated with the kernel function and determines the statistical importance of the data in the temporal vicinity of a time t. For instance, if the kernel function is chosen to be the probability density function (PDF) of the standard normal distribution,

$$W(s) = \frac{1}{\sqrt{(2\pi)}} \exp\left(-\frac{s^2}{2}\right), \tag{4}$$

then b is the standard deviation of the normal PDF $W\left(\frac{t-s}{b}\right)$. Hence, only data points within the window $[t-b, t+b]$ significantly contribute to the estimator in (2).

The effect of the Gaussian kernel on the estimation of $\mu(t)$ is exemplified on a time series generated via a persistent switching process between two processes, each wiggling around a constant mean value. The estimators $\hat{\mu}(t)$ for different choices of the window size b are depicted in Figure 1. As expected, although the estimators for the smallest window size give good local estimations of the respective constant mean, they are noisy and not constant. Moreover, the estimator becomes poor for time points close to the beginning or the end of the time series

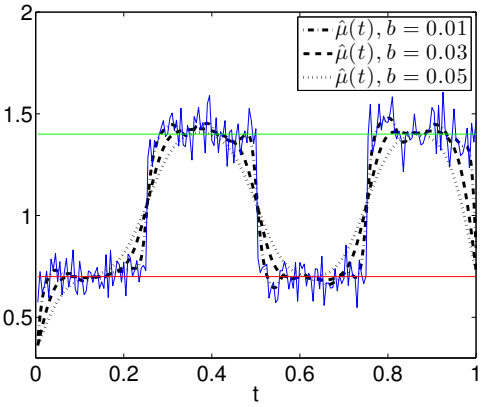


Figure 1. Illustration of the local kernel method on a time series generated via a persistent switching process between two processes each wiggling around a constant mean value ($\mu_1 = 0.7$, $\mu_2 = 1.4$). The estimator $\hat{\mu}(t)$ strongly depends on the specific choice of the window size. Results for a Gaussian kernel and $b = 0.01$, 0.03 and 0.05.

which is due to an insufficient statistics. In contrast, the graph of the estimators for the biggest window size is smooth but gives poorly local estimations and, hence, does not capture the intrinsic dynamics of the time series. Consequently, choosing the "right" window size b is an *ill-posed* optimization problem which is basically due to the local ansatz of the approach and the danger of overfitting.

The approach presented herein can be understood as a method to adaptively identify nonlocal kernel-functions which enforces optimal regularization of the estimators. The basic underlying idea is to simultaneously detect the hidden switching process between persistent regimes (clusters) and their respective optimal parameters characterizing local substitute models [39; 38; 31; 32; 30; 33]. Mathematically, the hidden (affiliation) process defines a curve in parameter space. The optimal paths and the associated optimal parameters of the local models are characterized via the minimization of an appropriate *clustering functional* measuring the quality of data approximation in terms of a fixed number of local error measures. In order to avoid overfitting, or more generally spoken, to ensure *well-posedness* of the clustering problem as an inverse problem, the smoothness of paths as a function of time is limited in some appropriate function space, e.g., the Sobolev H^1 space [31; 33] or the larger class BV, consisting of functions with *bounded variations* [33].

The cluster algorithms arising from the H^1 approach and the BV-approach partially result from *finite element (FE) discretization* of the 1-dimensional cluster functional. This allows us to apply methods from the broad repository of existing FE methods from the numerics of partial differential equations (PDEs). The H^1-smoothness of the paths in parameter space is indirectly enforced by a *Tikhonov regularization* leading to numerically expensive constrained quadratic minimization problems during the course of minimization of the cluster functional. In contrast, the variational formulation in the BV-space amounts to solving linear programming problems with linear constraints and, most importantly, allows the direct control of the regularization of the paths in parameter space. The entire FEM-BV approach will be explained in detail in Section 2.

The FEM-BV approach has two advantages; We neither have to make any assumptions *a priori* on the probabilistic nature of the data, i.e., on the underlying distribution of the data, nor we have to assume *stationarity* for the analysis of the time series (in contrast to standard methods such as HMMs, GMMs or local kernel methods). Moreover, as demonstrated in [31], the method covers geometrical cluster approaches as well as dynamical ones. Furthermore, we will discuss in Section 2.h the relation of the proposed approach to probabilistic methods.

The outcome of the FEM-BV methodology depends on the prescribed number of clusters (local models) as well as on the prescribed regularity. Hence, the optimal choice of these parameters is crucial for the interpretation and the meaningfulness of the analysis. The new idea presented in this paper is to select the model

that describes the data best while involving the least number of free parameters by combining an information theoretic measure — Akaike's information criterion (AIC) [1] — with the maximum entropy approach [43; 44]. The resulting modified AIC then allows us to identify in a postprocessing step the optimal nonstationary data-based substitute model. The main advantage of the modified AIC approach (presented in this manuscript) to information theoretical approaches used until now is that no explicit assumptions on the parametric form of observables' distributions have to be made. The only assumption is that a scalar process describing the time-dependent error of the inverse problem is i.i.d.

Complementary to providing insight in the nonstationary behavior of the time series, the optimal substitute model lends itself for predicting the dynamics, e.g., for different initial values. The prediction, however, is restricted to time instances within the trained time span (as the underlying transition process in parameter space is only available for that span). To overcome that restriction, a substitute model for the (nonstationary) transition process itself is derived. Combining the two data based models leads to a self-contained model that allows us to predict the dynamics for any initial value at any time instance.

1.a. ***New contributions and organization of the paper.*** The main purpose of this manuscript is threefold. First, in Section 2 we provide a complete, unified and simplified derivation of the FEM-BV methodology originally introduced in [29; 30; 31; 32; 33; 34; 35] for analyzing nonstationary time series. Thereby, we exemplify in Section 2.c the derivation of the framework for different models to give a guideline how the developed methodology can be adapted and redesigned for new applications. For the first time, specifically, we adapt the FEM-BV approach to: (i) analyze periodic and partially observed (projected) data (torsion angles of a biomolecule) and (ii) to pattern recognition in discrete data sequences (first chromosome of the yeast).

The second purpose is to close the gap between the FEM-BV approach and classical methods by investigating the assumptions and conditions under which the FEM-BV methodology reduces to well-known methods for analyzing (non-)stationary time series. For details see Section 2.h. Particularly, for the first time we clarify in Section 2.g under what conditions the solution of the variational problem (associated with the interpolation of the inverse model) can be interpreted as a direct interpolation model (mixture model) for the data under consideration.

Additionally, we present a unified strategy for model selection in Section 3 that allows the selection of an optimal mixture model — optimal in the sense that the model provides maximal meaningfulness under minimal assumptions on the data. The new model selection criterion combines a well known information criterion with the maximum entropy approach for the inference of probabilistic distributions from observables without assuming any parametric form.

All these three aspects are eventually combined in a self-contained scheme for predicting the nonstationary dynamics of the data beyond the analyzed time horizon. The prediction scheme is motivated and described in detail in Section 4.

Finally, the applicability and usefulness of the presented methods is demonstrated in Section 5 by analyzing realistic data ranging from torsion angle time series of a biomolecule (trialanine), DNA nucleotide sequence data (from the first chromosome of the yeast *Saccharomyces cerevisiae*) and financial data (prices of oil futures). We end this manuscript by giving a conclusion in Section 6.

2. Finite element clustering method

2.a. ***The model distance function.*** Modeling processes in real world applications amounts to seeking an appropriate parametric model function which is considered to govern (explain) well the observed process. Suppose the observable of interest, denoted by x_t, is a d-dimensional vector. Furthermore, without loss of generality, assume that the time series of observations is given at times $t = 0, 1, \ldots, T$. Then, the *direct mathematical model* is a function $f(\cdot)$ that relates an observation $x_t \in \Psi \subset \mathbb{R}^d$ at a time $t \geq 0$ to the history of observations up to the time t and a time-dependent set of parameters $\theta(t)$ from some parameter space Ω. Formally, the relation is written as[1]

$$x_t = f(x_t, \ldots, x_{t-m}, \theta(t)) \quad t \geq m, \tag{5}$$

where $m \geq 0$ is the memory depth of the history dependence. Notice that the formulation in (5) is most general in that it also covers implicit dependencies. See, e.g., (26) in Section 2.c.ii.

The model function can be deterministic or can denote a random process. For instance, the simplest model function incorporating randomness is given by

$$x_t = f(\theta(t)) \stackrel{\text{def}}{=} \theta(t) + \varepsilon_t, \tag{6}$$

where $\{\varepsilon_t\},\ t \geq 0$ is a family of i.i.d. random variables with $\mathbb{E}[\varepsilon_t] = 0,\ t \geq 0$. The random variables ε_t model, for instance, errors in the measurement of observables or they capture unresolved scales of a physical process such as fast degrees of freedoms. Thus, the model function in (6) corresponds to the assumption that the process under consideration has no dynamics and no memory.

Suppose we knew the parameters m and $\theta(t),\ t \geq 0$ then the *direct mathematical problem* would be to find a process $x_t,\ t \geq 0$ satisfying the direct model in (5). Here we are interested in the opposite question. Suppose we are given a time series of observations $\boldsymbol{X} = (x_t), t = 0, \ldots, T$ and a known memory depth m. What are the optimal parameters, i.e., the parameter function $\theta^*(t)$ explaining the given time

[1]For notational convenience, we prefer (5) to the equivalent relation $0 = F(x_t, \ldots, x_{t-m}, \theta(t))$.

series of observations best? This *inverse problem* makes only sense if "best" is quantified in terms of a *fitness function* measuring the quality of the approximation for a given set of parameters. Throughout this manuscript a fitness function is denoted by

$$g(x_t, \dots, x_{t-m}, \theta(t)) : \Psi^{m+1} \times \Omega \mapsto \mathbb{R}. \tag{7}$$

Particularly, any metric $d(\cdot, \cdot) : \Psi \times \Psi \mapsto \mathbb{R}_0^+$ on Ψ naturally induces a fitness function by defining $g(\cdot)$ as

$$g(x_t, \dots, x_{t-m}, \theta(t)) = \left(d\left(x_t, \mathbb{E}\left[f(x_t, \dots, x_{t-m}, \theta(t))\right]\right)\right)^2. \tag{8}$$

For instance, a reasonable model distance function for the direct mathematical model in (6) is induced by the Euclidean norm, i.e.,

$$g(x_t, \theta(t)) = \|x_t - \theta(t)\|_2^2. \tag{9}$$

By employing a metric, the resulting function $g(\cdot)$ measures the model error as the squared distance between x_t and the output of the *average* model function. Therefore, we call $g(\cdot)$ *model distance function* rather than fitness function. However, any function g that is bounded from below measuring the approximation quality is admissible within the following variational framework.

With the model distance function at hand, the optimal parameters explaining the time series "best" can now formally be characterized as those satisfying the *variational problem*

$$\boldsymbol{L} \stackrel{\text{def}}{=} \sum_{t=m}^{T} g(x_t, \dots, x_{t-m}, \theta(t)) \to \min_{\theta(t) \in \Omega}. \tag{10}$$

From now on, we will refer to $\boldsymbol{L}$ as the *model distance function*. In general, the variational problem in (10) is *ill-posed* in the sense of Hadamard [26] as the parameter space Ω might be high- or even infinite-dimensional and, hence, may lead to underdetermined or trivial solutions. For instance, the variational problem associated with the model distance function in (9) admits the trivial but meaningless solution (e.g., regarding the prediction skill of such a model, it requires the exact knowledge of the infinite-dimensional function x_t at all times)

$$\theta^*(t) = x_t, \quad t = 0, \dots, T. \tag{11}$$

In order to avoid such trivial solutions, the variational problem needs to be regularized.

The key idea of an appropriate regularization is based on the observation that in many real world processes the parameter function $\theta(t)$ varies much slower than the observable x_t in itself. Hence, *local stationarity* of the parameter function $\theta(t)$ is a reasonable assumption, which eventually helps to overcome the ill-posedness

of the variational problem in (10). Formally, we assume the existence of $\boldsymbol{K}$ different stationary yet unknown parameters $\Theta = (\theta_1, \ldots, \theta_{\boldsymbol{K}})$ and time-dependent weights $\Gamma(t) = (\gamma_1(t), \ldots, \gamma_{\boldsymbol{K}}(t))$ such that the model distance function $g(\cdot)$ can be expressed as a linear combination of *local* model distance functions, i.e.,

$$g(x_t, \ldots, x_{t-m}, \theta(t)) = \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) g(x_t, \ldots, x_{t-m}, \theta_i), \tag{12}$$

with $(\gamma_1(t), \ldots, \gamma_{\boldsymbol{K}}(t))$ satisfying the convexity constraints

$$\begin{aligned} \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) = 1, \quad \forall t, \\ \gamma_i(t) \geq 0, \quad \forall t, i. \end{aligned} \tag{13}$$

We call the vector $\Gamma(t)$ affiliation vector and we will use the shorthand $\Gamma = (\Gamma(t))_{t=m,\ldots,T}$. It is important to realize that, unlike in standard methods such as GMM/HMM, we do not assume the existence of $\boldsymbol{K}$ different local stationary models. Our assumption is more general since it is an assumption on the decomposability of the model error. However, as indicated by the name "affiliation vector", under certain conditions the entries of $\Gamma(t)$ can be interpreted as weights in a mixture model of local models (Section 2.g).

Inserting the interpolation ansatz (12) into the model distance function yields the *average cluster functional*

$$\boldsymbol{L}(\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma) = \sum_{t=m}^{T} \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) g(x_t, \ldots, x_{t-m}, \theta_i), \tag{14}$$

which is the key-object in the FEM-BV methodology. Additionally to the optimal (stationary) parameters $\Theta^* = (\theta_1^*, \ldots, \theta_{\boldsymbol{K}}^*)$ we seek for the optimal affiliation vectors Γ^*, which are finally characterized by the regularized variational problem

$$\boldsymbol{L}(\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma) \to \min_{\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma} \tag{15}$$

with Γ subject to the constraints in (13).

2.b. ***Numerical solution of the variational problem via the subspace algorithm.*** Even for the regularized variational problem derived from the simple model given in (6) there does not exist any analytical expression for the *global minimizer*, which is due to the nonlinearity of the average cluster functional and the convexity constraint on Γ. Fortunately, for many cases the model distance function $g(\cdot)$ is *convex* and analytical expressions for the unique optimal parameter Θ^* are available provided that Γ is given and *fixed*. The same holds true for the optimal Γ^* if the parameters Θ are fixed. Under weak conditions on the model distance function

Require: Time series $\boldsymbol{X}$, number of clusters $\boldsymbol{K}$, persistence $\boldsymbol{C}$, initial affiliations Γ^0.
Ensure: Locally optimal affiliations Γ^*, optimal parameters Θ^*.

Repeat until convergence

(1) Compute $\Theta^{(s+1)}$ for fixed $\Gamma^{(s)}$ via the unconstrained minimization problem

$$\Theta^{(s+1)} = \underset{\Theta}{\operatorname{argmin}}\, \boldsymbol{L}(\Theta, \Gamma^{(s)}) \tag{16}$$

(2) Compute $\Gamma^{(s+1)}$ for fixed $\Theta^{(s+1)}$ via the constrained minimization problem

$$\Gamma^{(s+1)} = \underset{\Gamma}{\operatorname{argmin}}\, \boldsymbol{L}(\Theta^{(s+1)}, \Gamma) \tag{17}$$

subject to (13).

Algorithm 1. The subspace algorithm.

$g(\cdot)$ it was proven in [31] that iterating over these two steps yields an algorithm guaranteed to converge to a *local minimum* of the average cluster functional $\boldsymbol{L}$.

Throughout this paper when we speak of the *subspace algorithm*, we are actually referring to an implementation of the iterative scheme described above and formally summarized in Algorithm 1.

The subspace algorithm converges only to a local minimum. In order to find the global minimum, an annealing-like Monte Carlo strategy can be employed, i.e., the iterative procedure is started over several times with randomly initialized $\Gamma^{(0)}$. If the number of repetitions is sufficiently large then the best solution among the local minimizer is (almost sure) the global minimizer Γ^* and Θ^*. Notice that the described strategy for finding the global minimizers can straightforwardly be parallelized.

2.c. ***Four important models.*** In this section we introduce four important models that are broadly used in time series analysis and we derive their respective associated variational formulations. Numerical results will be given in Section 5.

2.c.i. *Model I: Geometrical clustering.* In the Section 2.a we introduced the simplest nontrivial model one can think of; a model without memory,

$$x_t = \theta(t) + \varepsilon_t, \tag{18}$$

where $x_t \in \mathbb{R}^d$ and ε_t denotes a noise process. If we choose the model distance function induced by the Euclidean norm,

$$g(x_t, \theta(t)) = \|x_t - \theta(t)\|_2^2, \tag{19}$$

then the regularized minimization problem in (15) simplifies to

$$\boldsymbol{L}(\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma) = \sum_{t=0}^{T} \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) \|x_t - \theta_i\|_2^2 \to \min_{\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma} \tag{20}$$

subject to the constraints in (13). For fixed Γ the optimal $\Theta^* = (\theta_1^*, \ldots, \theta_{\boldsymbol{K}}^*)$ takes the form [31]

$$\theta_i^* = \frac{\sum_{t=0}^{T} \gamma_i(t) x_t}{\sum_{t=0}^{T} \gamma_i(t)}. \tag{21}$$

Furthermore, for fixed Θ the optimal affiliations are given by [33]

$$\gamma_i^*(t) = \begin{cases} 1 & \text{if } i = \operatorname{argmin}_j g(x_t, \theta_j) = \operatorname{argmin}_j \left\{ \left\| x_t - \theta_j \right\|_2^2 \right\}, \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

which readily follows from the convexity conditions in (13).

The resulting subspace algorithm has a very nice interpretation: it is the well-known and popular k-means algorithm for clustering geometrical data. To see that note that each affiliation vector is deterministic, i.e., exactly one component is 1.0 while the remaining ones are 0.0. If we define the set $S_i = \{x_t : \gamma_i^*(t) = 1\}$ then, by definition

$$\|x_t - \theta_i\|_2 \leq \left\| x_t - \theta_j \right\|_2 \quad \forall x_t \in S_i, \ j = 1, \ldots, \boldsymbol{K}, \tag{23}$$

and the optimal θ_i^* reduces to the centroid of the point set S_i,

$$\theta_i^* = \frac{1}{|S_i|} \sum_{x_t \in S_i} x_t. \tag{24}$$

2.c.ii. *Model II: Takens-PCA clustering.* A prominent example of a memoryless model exhibiting dynamics is motivated by the observation that in many applications the essential dynamics of a high-dimensional process can be approximated by a process on low-dimensional manifolds without significant loss of information [70]. Recently, several cluster methods have been introduced which are based on the decomposition of time series according to their *essential linear attractive manifolds*, allowing the analysis of data of very high dimensionality with low-dimensional dynamics [40; 29; 39; 38].

Formally, assume that the linear submanifolds are spanned by $Q(t) \in \mathbb{R}^{d \times n}$ consisting of $n \ll d$ orthonormal d-dimensional vectors, i.e., $Q^\dagger(t) Q(t) = \mathrm{Id}_n$ where Id_n denotes the n-dimensional identity matrix. To motivate the following direct mathematical model, suppose that x_t lives on the linear subspace spanned by $Q(t)$. Orthonormality then implies

$$x_t = Q(t) Q^\dagger(t) x_t, \tag{25}$$

where $Q(t)Q^{\dagger}(t)$ is the orthogonal projector on the linear subspace at time t. However, in applications we only have $x_t \approx Q(t)Q^{\dagger}(t)x_t$, which leads to the general model function

$$(x_t - \mu_t) = Q(t)Q^{\dagger}(t)(x_t - \mu_t) + \varepsilon_t, \tag{26}$$

where the center vector $\mu_t \in \mathbb{R}^d$ is the affine translation of the linear subspace and ε_t is again some noise process with $\mathbb{E}[\varepsilon_t] = 0$. As shown in [38], adopting the model distance function ($\theta(t) = (\mu(t), Q(t))$)

$$g(x_t, \theta(t)) = \left\| (x_t - \mu_t) - Q(t)Q^{\dagger}(t)(x_t - \mu_t) \right\|_2^2 \tag{27}$$

results in analytical closed expressions for the optimal parameters. The center vectors $\mu_i^* \in \mathbb{R}^d$ are given by

$$\mu_i^* = \frac{\sum_{t=0}^{T} \gamma_i(t) x_t}{\sum_{t=0}^{T} \gamma_i(t)} \tag{28}$$

and the optimal matrices Q_i^* satisfy an eigenvalue problem, respectively,

$$\left(\sum_{t=0}^{T} \gamma_i(t)(x_t - \mu_i)(x_t - \mu_i)^{\dagger} \right) Q_i^* = Q_i^* \Lambda_i. \tag{29}$$

For fixed Θ, the optimal Γ^* is given analogously by (22).

2.c.iii. *Model III: Discrete (or categorical) model.* An alternative technique to capture the essential dynamics of a complex system is *coarse graining* of the process under consideration. The coarse grained process is a *discrete* process, i.e., it attains only values in a finite set of discrete objects. Prominent examples are, e.g., conformational dynamics of (bio-)molecules [67] or climate research [30].

Let $\boldsymbol{X} = (x_1, \ldots, x_T)$ be a discrete time series and without loss of generality we denote the discrete state space as $\mathcal{S} = \{1, \ldots, M\}$. In order to apply the variational framework we have to specify a model function and an appropriate model distance function that are not readily available due to the discreteness of the state space. Instead of considering the original data, the key idea here is to uniquely identify each datum x_t with a discrete probability distribution π_t. More precisely, we define $\pi_t = (\pi_t(1), \ldots, \pi_t(M))$ as the discrete Dirac measure with respect to $x_t \in \mathcal{S} = \{1, \ldots, M\}$,

$$\pi_t(s) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } s = x_t, \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$

Viewing discrete distributions as real valued vectors allows us to make use of, e.g., the model function given in (6), here written as

$$\pi_t = \theta_t + \varepsilon_t, \tag{31}$$

subject to the constraint that $\theta_t = (\theta_t(1), \ldots, \theta_t(M))$ is a discrete probability distribution,

$$\theta_t(s) \geq 0 \quad \text{and} \quad \sum_{s=1}^{M} \theta_t(s) = 1. \tag{32}$$

Moreover, ε_t is a noise process as in the previous models.

Since we are particularly dealing with probability distributions, we define the model distance function by means of a metric tailored to respect the underlying probability space. Specifically, we chose the famous Kullback–Leibler divergence, also referred to as the relative entropy, defined as

$$d_{KL}(\mu, \eta) = \sum_{s \in \mathcal{S}} \mu(s) \log \frac{\mu(s)}{\eta(s)} \tag{33}$$

for any two discrete probability distributions μ and η on the same probability space. For an overview of metrics and divergences on probability spaces see [25], for example.

The relative entropy directly induces a model distance function by defining

$$g(x_t, \theta_t) \stackrel{\text{def}}{=} g(\pi_t, \theta_t) \stackrel{\text{def}}{=} d_{KL}(\pi_t, \theta_t) = -\log \theta_t(x_t). \tag{34}$$

A short calculation shows that the regularized minimization problem

$$\boldsymbol{L}(\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma) = -\sum_{t=0}^{T} \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) \log \theta_i(x_t) \to \min_{\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma} \tag{35}$$

subject to the constraints (13) and (32) admits analytical solutions; the optimal discrete probability distribution $(\theta_1^*, \ldots, \theta_K^*)$ takes the form

$$\theta_i^*(s) = \frac{\alpha_{i,s}}{\sum_{z \in \mathcal{S}} \alpha_{i,z}} \quad \text{with } \alpha_{i,s} = \sum_{t=0}^{T} \delta_{x_t,s} \gamma_i(t), \quad s = 1, \ldots, M \tag{36}$$

and the optimal affiliation function Γ^* is given analogously by (22).

2.c.iv. *Model IV: Markov regression model.* The strategy proposed in Section 2.c.iii to analyze time series of discrete observations can loosely be described as geometrical clustering of probability distributions, geometrical in the sense that neither dynamics nor memory are assumed to be of importance.

A discrete probabilistic model including memory and dynamics is the famous Markov model. Generally, a discrete Markov process describes the evolution of a transition process between a finite number of discrete states by means of time-dependent one-step transition probabilities. If the transition probabilities are stationary (time-homogeneous) then the process is called a Markov chain and it is one of the most exploited families of processes in this class of probabilities models.

Formally, a stationary Markov process x_t on a discrete state space $\mathcal{S} = \{1, \ldots, M\}$ is uniquely characterized by a time-independent transition (stochastic) matrix $P \in \mathbb{R}^{M \times M}$ (comprising of the stationary one-step transition probabilities) and an initial distribution $\pi_0 \in \mathbb{R}^M$. The evolution of the state probability vector $p(t) \in \mathbb{R}^M$, defined as

$$p_j(t) \stackrel{\text{def}}{=} \mathbb{P}[x_t = j], \quad j \in \mathcal{S}, \tag{37}$$

is then governed by the *master equation*,

$$p^\dagger(t+1) = p^\dagger(t)P, \quad t = 0, 1, 2, \ldots, T-1. \tag{38}$$

For more details on Markov chains, we refer the interested reader to, e.g., [9].

Recently in [34], the opposite question was addressed: suppose we are given a time series of *probability distributions* (π_t), $\pi_t \in \mathbb{R}^M$, $t = 0, 1, \ldots, T$ and, additionally, a series of external data $u(t) \in \mathbb{R}^k$. What is an appropriate *nonstationary Markov regression model* explaining the given time series of distributions conditioned on the external factors best? Following the lines of the FEM-BV approach and motivated by the stochastic master Equation (38), it is reasonable to consider the direct model function

$$\pi_{t+1}^\dagger = \pi_t^\dagger P(t, u(t)) + \varepsilon_t \tag{39}$$

where ε_t is a noise process as in the previous models and $P(t, u(t)) \in \mathbb{R}^{M \times M}$ is stochastic, i.e.,

$$\{P(t, u(t))\}_{vw} \geq 0 \quad \forall v, w, t, u(t), \tag{40}$$

$$P(t, u(t))\mathbf{1}_M = \mathbf{1}_M \quad \forall t, u_t \tag{41}$$

with $\mathbf{1}_M = (1, \ldots, 1) \in \mathbb{R}^M$.

Additional to depending on the (resolved) external factors $u(t) \in \mathbb{R}^k$, notice that the transition matrices may explicitly depend on the time t. For details see [34].

The interpolation of the model distance function

$$g\left(\pi_{t+1}, \pi_t, P(t, u(t))\right) = \left\| \pi_{t+1}^\dagger - \pi_t^\dagger P(t, u(t)) \right\|_2^2 \tag{42}$$

results in

$$g\left(\cdot, \cdot, P(t, u(t))\right) = \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) g\left(\cdot, \cdot, P^{(i)}(u(t))\right) \tag{43}$$

where the stationary transition matrices (parameters), $P^{(i)}(u(t)) \in \mathbb{R}^{M \times M}$ $i = 1, \ldots, \boldsymbol{K}$, have the form

$$P^{(i)}(u(t)) = P_0^{(i)} + \sum_{l=1}^{k} u_l(t) P_l^{(i)} \quad i = 1, \ldots, \boldsymbol{K} \tag{44}$$

with $P_0^{(i)}$, $P_l^{(i)} \in \mathbb{R}^{M\times M}$ $i = 1, \ldots, \boldsymbol{K}$ satisfying the constraints

$$P_0^{(i)} \geq 0 \quad \text{(elementwise)}, \tag{45}$$

$$P_0^{(i)}\mathbf{1}_M = \mathbf{1}_M, \tag{46}$$

$$P_l^{(i)}\mathbf{1}_M = 0 \quad l = 1, \ldots, k. \tag{47}$$

Notice that the constraints (45)–(47) imply $P^{(i)}(u(t))\mathbf{1}_M = \mathbf{1}_M$ independently of $u(t)$. The elementwise nonnegativity is ensured by the constraints

$$P^{(i)}(u(t)) \geq 0 \quad i = 1, \ldots, \boldsymbol{K}, \ \forall u(t), \tag{48}$$

which explicitly involve the external data $u(t)$.

Assembling the pieces together, we finally end up with the variational problem

$$\boldsymbol{L}(\Theta, \Gamma) = \sum_{t=0}^{T-1}\sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) g\left(\pi_{t+1}, \pi_t, P_0^{(i)} + \sum_{l=1}^{k} u_l(t) P_l^{(i)}\right) \to \min_{\Theta,\Gamma} \tag{49}$$

subject to the constraints (45)–(48). Unfortunately, no analytical expressions exist for the optimal parameters due to the imposed constraints. Numerically, however, the optimal Markov regression models $P^{(i)}(t, u(t))$ are given by solutions of $\boldsymbol{K}$ independent constrained quadratic programs. For the convenience of the reader, they are stated in an Appendix.

The main challenge in numerical computation of the optimal parameters lies in the enforcement of the constraints in (48) as a linear increase in the number of external factors causes an exponentially increase in time and memory for minimizing (49). As shown in [34], the computational time and memory consumption can be reduced by exploiting that (48) attains its unique maximum/minimum in a corner of the convex hull of the set $\{u(t) : t = 0, \ldots, T\}$. Hence, it is sufficient to requiring the constraints in (48) only for the corners. For example, if the convex hull is given by an k-dimensional hypercube then the reduced number of constraints, 2^k, is independent of the length of the time series. This allows to substitute the time-dependent set of constraints (48) by a time-independent set, making the entire optimization problem numerically tractable.

2.d. ***Regularization of*** $\boldsymbol{\Gamma}$**.** As indicated in the examples introduced in the previous section, for given parameters Θ the optimal Γ^* is given in terms of the model distance function (compare (22), for example),

$$\gamma_i^*(t) = \begin{cases} 1 & \text{if } i = \operatorname{argmin}_j \{g(x_t, \ldots, x_{t-m}, \theta_j)\}, \\ 0 & \text{otherwise}, \end{cases} \tag{50}$$

where each datum x_t, $t \geq m$ is uniquely (deterministically) assigned to a single cluster. However, even for the global optimal parameters Θ^*, the resulting optimal

Γ^* might be a highly nonregular function. For instance, Γ^* might rapidly oscillate between the $\boldsymbol{K}$ different clusters rather than describing a smooth and persistent transition process. In other words, the optimal Γ^* does not *continuously* depend on the data, which is again a violation of Hadamard's postulate of a *well-posed* problem. Consequently, the variational problem has to be regularized again.

One approach is to first incorporate some *additional information* about the regularity of the observed process by restricting the time dependent function $\Gamma(\cdot)$ on an appropriate function space and then apply a finite Galerkin discretization of this infinite-dimensional Hilbert space. In the context of Tikhonov-based FEM-BV methodology, this was done by restricting the functions $\gamma_i(\cdot)$ on the function space of *weakly differentiable* functions. One way to incorporate this *a priori information* into the optimization is to modify the variational problem in (15) by writing it in the *Tikhonov-regularized* form [31]

$$\boldsymbol{L}^{\varepsilon}(\Theta, \Gamma, \varepsilon^2) \stackrel{\text{def}}{=} \boldsymbol{L}(\Theta, \Gamma) + \varepsilon^2 \sum_{i=1}^{\boldsymbol{K}} \|\partial_t \gamma_i\|^2_{L_2(0,T)} \to \min_{\gamma_1,\ldots,\gamma_{\boldsymbol{K}} \in H^1(0,T),\Theta}, \quad (51)$$

where the norm $\|\partial_t \gamma_i\|^2_{L_2(0,T)} = \int_0^T (\partial_t \gamma_i(t))^2 \mathrm{d}t$ measures the smoothness of the function $\gamma_i(\cdot)$. A similar form of penalized regularization was first introduced by A. Tikhonov to solve ill-posed linear least-squares problems [71] and has been frequently used for nonlinear regression analysis in the context of statistics [28] and multivariate spline interpolation [74].

The main problem one faces in this approach is the lack of the direct control of the persistence of γ_i. To be more precise, Tikhonov regularization does not allow us to directly incorporate the desired *persistence constraints*

$$\|\partial_t \gamma_i\|^2_{L_2(0,T)} \le \boldsymbol{C}, \quad i = 1, \ldots, \boldsymbol{K}, \quad (52)$$

where $0 \le \boldsymbol{C}$ bounds the smoothness of the functions $\gamma_i(\cdot)$. Another disadvantage of the H^1 approach is the exclusion of functions with discontinuities such as jumps, which is due to the requirement of weak differentiability. Fortunately, the two problems can be overcome by considering a larger function space.

2.e. ***Persistence in the BV sense.*** A less restrictive class of functions is the class of functions with *bounded variation* $BV([0, T])$, consisting of functions $f : [0, T] \to \mathbb{R}$ with

$$\|f\|_{BV} = \sup_{0=t_0<t_1<\ldots<t_M=T} \left\{ \sum_{i=0}^{M-1} |f(t_{i+1}) - f(t_i)| \right\} < \infty, \quad (53)$$

where the supremum is taken over all partitions of the interval $[0, T]$. Notice that in the time-continuous case $H^1(0, T) \subset BV(0, T)$ holds true (cf. [58]), so "smooth" H^1-transitions between cluster states are not excluded. However, the BV-norm of

a function does not require any notion of differentiability and the class $BV[0, T]$ covers transition processes with jumps between clusters.

For the remainder of this section, the memory depth m is, without loss of generality, assumed to be zero. In the following, we consider the functions $\gamma_i,\ i = 1, \ldots, \boldsymbol{K}$ as *discrete* functions (vectors), which is emphasized by denoting $\gamma_i \in \mathbb{R}^{T+1}$. Now we are prepared to formulate the persistence condition in the time-discrete BV sense:

$$\|\gamma_i\|_{BV} = \sum_{t=0}^{T-1} |\gamma_i(t+1) - \gamma_i(t)| \leq \boldsymbol{C}, \quad i = 1, \ldots, \boldsymbol{K}, \tag{54}$$

where $0 \leq \boldsymbol{C}$ is an upper bound for the maximal number of transitions between the cluster state i and the remaining ones. In the rest of this section we will show that the additional BV-constraints lead to a numerically convenient characterization of Γ via a *linear minimization problem with linear constraints*.

To this end, for given $\Theta = (\theta_1, \ldots, \theta_{\boldsymbol{K}})$ we define the *row vectors*

$$g_{\theta_i} = (g(x_0, \theta_i), \ldots, g(x_T, \theta_i)) \in \mathbb{R}^{T+1}, \tag{55}$$

$$\gamma_i = (\gamma_i(0), \ldots, \gamma_i(T)) \in \mathbb{R}^{T+1}. \tag{56}$$

Then, the variational problem in (15) transforms to

$$\boldsymbol{L}(\theta_1, \ldots, \theta_{\boldsymbol{K}}, \Gamma) = \sum_{i=1}^{\boldsymbol{K}} \langle \gamma_i, g_{\theta_i} \rangle_2 \to \min_{\Gamma, \Theta}, \tag{57}$$

subject to the constraints

$$\|\gamma_i\|_{BV} \leq \boldsymbol{C} \quad i = 1, \ldots, \boldsymbol{K}, \tag{58}$$

$$\sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) = 1 \quad t = 0, \ldots, T, \tag{59}$$

$$\gamma_i(t) \geq 0 \quad t = 0, \ldots, T,\ i = 1, \ldots, \boldsymbol{K}. \tag{60}$$

Unfortunately, the additional constraints (58) turn the variational problem in (57) into a *nondifferentiable* one. As a remedy, we retransform the problem into a differentiable one by applying an upper-bound technique.

Suppose we had $\eta_i(0), \ldots, \eta_i(T-1) \in \mathbb{R}$ satisfying the constraints

$$|\gamma_i(t+1) - \gamma_i(t)| \leq \eta_i(t) \quad t = 0, \ldots, T-1, \tag{61}$$

$$\sum_{t=0}^{T-1} \eta_i(t) \leq \boldsymbol{C}, \tag{62}$$

$$\eta_i(t) \geq 0 \qquad t = 0, \ldots, T-1, \tag{63}$$

then γ_i would satisfy the BV-constraint in (58). The key observation is that (61) holds true for $t \geq 0$ if and only if the following two *linear* inequalities hold true:

$$\gamma_i(t+1) - \gamma_i(t) - \eta_i(t) \leq 0, \tag{64}$$

$$-\gamma_i(t+1) + \gamma_i(t) - \eta_i(t) \leq 0. \tag{65}$$

Consequently, if the upper bounds $\eta_i = (\eta(0), \ldots, \eta(T-1))$ are considered as additional unknowns (additional to the unknowns γ_i), then the BV-constraint in (58) is satisfied if and only if the linear constraints (62)–(65) are satisfied.

Notice that the constraints (59)–(60) are *linear* constraints too. Finally, by defining

$$\omega = (\gamma_1, \ldots, \gamma_{\boldsymbol{K}}, \eta_1, \ldots, \eta_{\boldsymbol{K}}) \in \mathbb{R}^{\boldsymbol{K}(2T+1)}, \tag{66}$$

$$c(\Theta) = (g_{\theta_1}, \ldots, g_{\theta_{\boldsymbol{K}}}, \underbrace{0, \ldots, 0}_{\boldsymbol{K}T \text{ times}}) \in \mathbb{R}^{\boldsymbol{K}(2T+1)} \tag{67}$$

we can express the original *nondifferentiable* optimization problem (57)–(60) as the following *differentiable* optimization problem,

$$\langle c(\Theta), \omega \rangle_2 \to \min_{\omega, \Theta} \tag{68}$$

subject to

$$\begin{aligned} A_{\text{eq}}\omega &= b_{\text{eq}}, \\ A_{\text{neq}}\omega &\leq b_{\text{neq}}, \\ \omega &\geq 0, \end{aligned} \tag{69}$$

where A_{eq} and b_{eq} readily result from the constraints (59) and A_{neq} and b_{neq} from (60) and ((62)–(65)).

The solution of the above minimization problem can be approached via the subspace iteration procedure presented in Section 2.b. Particularly, for fixed Θ the problem reduces to a standard *linear program*, which can efficiently be solved by standard methods such as the Simplex method or interior point method. Completely analogously to the Tikhonov-regularized FEM-BV methodology [31], it can be demonstrated that the iterative procedure converges towards a local minimum of the problem (68)–(69) if some appropriate assumptions (convexity and differentiability) of the model distance function (8) are fulfilled.

Unfortunately, since the dimensionality of the variable ω scales as $\boldsymbol{K}(2T+1)$ the numerical solution of the problem (68)–(69) for a fixed value of Θ becomes increasingly expensive for long time series. Therefore a Finite Element Method (FEM) will be introduced in the next section to reduce the dimensionality of the above problem in a robust and controllable numerical manner.

2.f. ***FEM discretization.*** Solving the problem (68)–(69) is numerically expensive or even practically impossible for long time series, in terms of computational time as well as in terms of memory usage. To overcome these limitations, a FEM is proposed to reduce the dimensionality of the problem.

The idea is to approximate the (unknown) discrete functions $\gamma_i(t)$ by a linear combination of $N \ll T+1$ continuous functions $\{f_1(t), f_2(t), \ldots, f_N(t)\}$ with bounded variation, i.e.,

$$\gamma_i(t) = \sum_{j=1}^{N} \alpha_{ij} f_j(t) \quad t = 0, \ldots, T+1. \tag{70}$$

Traditionally, the finite element functions $f_j(t) \in BV[0,T]$ are defined as nonconstant functions on overlapping supports. For practical examples of standard finite element functions see, e.g., [8]. Here, however, we approximate the functions γ_i with *constant* ansatz functions defined on *nonoverlapping* supports. This approach is justified by the fundamental assumption that the time series under consideration is *persistent*.

Let $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_N = T$ be a partition dividing $[0,T]$ into N bins $[\tau_j, \tau_{j+1}]$, $j = 0, \ldots, N-1$ with $\tau_j \notin \mathbb{N}$, $j = 1, \ldots, N-1$ and assume that all the γ_i are piecewise constant on each of the intervals $[\tau_j, \tau_{j+1}]$. Moreover, let $\hat{\gamma}_i(j)$ denote the value of γ_i on $[\tau_j, \tau_{j+1}]$ and define

$$\hat{g}_{\theta_i}(j) \stackrel{\text{def}}{=} \sum_{t \in [\tau_j, \tau_{j+1}]} g_{\theta_i}(t). \tag{71}$$

Then, the variation problem in (57) reduces to

$$\boldsymbol{L}(\theta_1, \ldots, \theta_{\boldsymbol{K}}, \hat{\Gamma}) = \sum_{i=1}^{\boldsymbol{K}} \left\langle \hat{\gamma}_i, \hat{g}_{\theta_i} \right\rangle_2 \to \min_{\hat{\Gamma}, \Theta}, \tag{72}$$

with $\hat{\gamma}_i \in \mathbb{R}^N$, $\hat{g}_{\theta_i} \in \mathbb{R}^N$ and subject to the constraints

$$\|D\hat{\gamma}_i\|_1 \leq \boldsymbol{C} \quad i = 1, \ldots, \boldsymbol{K}, \tag{73}$$

$$\sum_{i=1}^{\boldsymbol{K}} \hat{\gamma}_i(t) = 1 \quad t = 0, \ldots, N-1, \tag{74}$$

$$\hat{\gamma}_i(t) \geq 0 \quad t = 0, \ldots, N-1,\ i = 1, \ldots, \boldsymbol{K}. \tag{75}$$

Analogously to the derivation given in the previous section, we finally end up with the FEM discretization (in the BV sense) of the original variational problem

in (15),

$$\left\langle \hat{c}(\Theta), \hat{\Gamma} \right\rangle_2 \to \min_{\hat{\Gamma}, \Theta} \tag{76}$$

subject to the linear constraints (73)–(75).

Notice that the number of unknowns has reduced to $\boldsymbol{K}(2N+1)$ being much less than $\boldsymbol{K}(2T+1)$ if $N \ll T$. Particularly, the number of unknowns and, hence, the number of constraints does not explicitly depend on the total length $T+1$ of the time series anymore. Hence, the final variational problem allows the analysis of long time series from real-world applications, as will be demonstrated in Section 5.

2.g. ***Identification of local models.*** The derivation of the average cluster functional is based on the assumption that the model distance at a fixed time t can be represented by a convex combination of model distances with respect to $\boldsymbol{K}$ stationary model parameters. Notice that this assumption is more general than the assumption of the existence of $\boldsymbol{K}$ *local stationary models*. Nevertheless, the identification of local stationary models gives additional insight into the data. More importantly, it allows the simulation and prediction of time series, which ultimately leads to constructing self-contained predictive models as will be explained in Section 4 below.

The identification of local stationary models depends crucially on the choice of the model distance function and the derived optimal affiliation function Γ^*. To see that, recall the formal interpolation ansatz in (12), i.e.,

$$g(x_t, \ldots, x_{t-m}, \theta(t)) = \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) g(x_t, \ldots, x_{t-m}, \theta_i). \tag{77}$$

Accordingly, if we could find an $\theta(t)$ such that (77) held true then the local model at time t would be given by $f(\,\cdot\,; \theta(t))$.

First suppose that the optimal Γ^* is deterministic, i.e., $\gamma_i^*(t) \in \{0, 1\}$. But this immediately implies

$$\theta(t) = \theta_i \quad \text{with } \gamma_i^*(t) = 1, \tag{78}$$

as the ansatz trivially holds true with that choice. In the case of a nondeterministic Γ^* the identification crucially depends on the model distance function. We exemplify that by considering the model distance function

$$g(x_t, \ldots, x_{t-m}, \theta(t)) = \left\| x_t - \mathbb{E}\left[f(x_t, \ldots, x_{t-m}, \theta(t)) \right] \right\|_2^2. \tag{79}$$

Theorem 2.1. *If the direct model function f is* linear *in θ then*

$$g\left(x_t, \ldots, x_{t-m}, \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t)\theta_i\right) \leq \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) g(x_t, \ldots, x_{t-m}, \theta_i) \tag{80}$$

The proof is straightforward and left for the interested reader. Consequently, if the interpolation on the right-hand side in (80) is small then the model distance function on the left-hand side with respect to $\theta(t) = \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t)\theta_i$ is small too. This, in turn, implies that the direct model function with respect to $\theta(t)$ is a good approximation for a local model function at time t.

The minimization of the average cluster functional justifies the notion

$$x_t \approx \hat{x}_t \stackrel{\text{def}}{=} \mathbb{E}\left[f(x_t, \ldots, x_{t-m}, \sum_{i=1}^{\boldsymbol{K}} \gamma_i^*(t)\theta_i^*)\right]. \tag{81}$$

However, the identification is only valid if the direct model function is linear with respect to its parameters and the model distance function is strict convex. This is the case for the model distance functions, e.g., in (19), (27), (34) and (42) described above.

2.h. ***Relation to classical methods of unsupervised learning.*** We have already seen that the direct model $x_t = \theta(t) + \varepsilon_t$ equipped with the model distance function $g(x_t, \theta(t)) = \|x_t - \theta(t)\|_2^2$ leads to the classical k-means algorithm for geometric clustering provided that no regularity condition ($\boldsymbol{C} = \infty$) is imposed on the affiliation function Γ (Section 2.c.i) and no FEM discretization is used for the numerical solution of the resulting variational problem. In this section we further clarify the link between the FEM-BV approach and classical methods for dynamical clustering. Particularly, we show that the presented method covers existent probabilistic approaches as special cases by choosing specific model distance functions and regularity constraints.

Let us first consider the discrete case, i.e., $x_t \in \mathcal{S} = \{1, \ldots, M\}$. A prominent approach for dynamical clustering of persistent discrete time series is the hidden Markov model [64]. Basically, it relies on three strong assumptions. Firstly, it is assumed that the hidden (persistent) process is governed by a time-homogeneous stationary Markov process. Secondly, it is assumed that an observation x_t (triggered by a jump of the hidden process) is distributed according to a stationary distribution conditional on the current hidden state. Finally, one has to assume that the observations are independent.

Here we make the most general assumption by imposing that the hidden process is nonstationary and non-Markovian. Specifically, we assume that an observation x_t is distributed according to a discrete distribution $\theta_i \in \mathbb{R}^{|S|}$ conditional on a hidden state $i \in \{1, \ldots, \boldsymbol{K}\}$, which in turn is drawn from a discrete distribution $\Gamma(t) \in \mathbb{R}^{\boldsymbol{K}}$.

Under the additional assumption of independence, the likelihood of a time series $\boldsymbol{X} = (x_t)$, $t = 0, \ldots, T$ takes the form

$$\mathcal{L}(\boldsymbol{X}; \Gamma, \Theta) = \prod_{t=0}^{T} \Big(\sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) \theta_i(x_t) \Big), \tag{82}$$

where we marginalize over the hidden states.

Theorem 2.2. *If the model distance function is defined as*

$$g(x_t, \theta_i) = -\log(\theta_i(x_t)) \tag{83}$$

then the associated average cluster functional is an upper bound of the negative log-likelihood,

$$-\log \mathcal{L}(\boldsymbol{X}; \Gamma, \Theta) \leq L(\Gamma, \Theta). \tag{84}$$

Proof. Notice that $-\log x$ is a convex function. Hence, by applying Jensen's inequality we conclude

$$\begin{aligned} -\log \mathcal{L}(\boldsymbol{X}; \Gamma, \Theta) &= -\sum_{t=0}^{T} \log \Big(\sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) \theta_i(x_t) \Big) \\ &\leq \sum_{t=0}^{T} \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) \left(-\log(\theta_i(x_t)) \right), \end{aligned} \tag{85}$$

where the upper bound in (85) is exactly *the average cluster functional* in (35) resulting from the reasoning in the third example in Section 2.c.iii. □

In the probabilistic approach, the optimal parameters (distributions) of the model are characterized by the ones that maximize the likelihood, i.e.,

$$(\Gamma^*, \Theta^*) = \operatorname*{argmax}_{\Gamma, \Theta} \mathcal{L}(\boldsymbol{X}; \Gamma, \Theta), \tag{86}$$

which is equivalent to minimizing the negative log-likelihood function,

$$(\Gamma^*, \Theta^*) = \operatorname*{argmin}_{\Gamma, \Theta} (-\log \mathcal{L}(\boldsymbol{X}; \Gamma, \Theta)). \tag{87}$$

Therefore, the minimizer of the average cluster functional in (35) can be considered as a good approximation of the maximizer Γ^*, Θ^* of the likelihood function in (82). The fundamental difference between the two approaches, however, is that in the FEM-BV approach non of the probabilistic assumptions on the nature of data have to be made in order to derive the average cluster functional (35).

The presented reasoning readily carries over to the continuous case, i.e., $x_t \in \mathbb{R}^d$, by defining the model distance function in terms of the assumed underlying

conditional probability density function $\rho(\,\cdot\,;\theta_i)$,

$$g(x_t,\theta_i) \stackrel{\text{def}}{=} -\log \rho(x_t;\theta_i). \tag{88}$$

It is straightforward to show that the upper bound for the negative log-likelihood associated with probabilistic model coincides with the average cluster functional resulting from the model distance function in (88).

For example, a widely used class of parametric probability density functions are the d-dimensional Gaussian distributions,

$$\rho_G(x_t;\mu_i,\Sigma_i) = ((2\pi)^d|\Sigma|)^{-1/2}\exp\!\big(-\tfrac{1}{2}(x_t-\mu_i)^\dagger\Sigma_i^{-1}(x_t-\mu_i)\big), \tag{89}$$

with mean $\mu_i \in \mathbb{R}^d$ and symmetric positive definite covariance matrix $\Sigma_i \in \mathbb{R}^{d\times d}$. The induced model distance function then reads

$$g(x_t,\mu_i,\Sigma_i) = \tfrac{1}{2}\big(\text{cst.} + \ln|\Sigma| + (x_t-\mu_i)^\dagger\Sigma_i^{-1}(x_t-\mu_i)\big). \tag{90}$$

Any method for inferring the optimal parameters of a Gaussian distribution relies specifically on the assumption that the data "lives" in the full d-dimensional space so that the covariance matrix is symmetric positive definite and, hence, invertible. Unfortunately, in many applications this assumption is not met because, e.g., the essential dynamics of a (Gaussian) process takes place in an n-dimensional submanifold with $n \ll d$. In the FEM-BV approach, this limitation can be circumvented by directly clustering with respect to the submanifolds by means of the PCA approach presented in Section 2.c.ii.

At the end of this section, we comment on the relation of the FEM-BV approach based on (90) to the stationary Gaussian mixture model (GMM). Analogously to the reasoning above, the negative log-likelihood associated with a GMM can be bounded from above, i.e.,

$$-\sum_t \log\!\Big(\sum_{i=1}^{\boldsymbol{K}} a_i\rho_G(x_t;\mu_i,\Sigma_i)\Big) \le -\sum_t\sum_{i=1}^{\boldsymbol{K}} a_i\log\rho_G(x_t;\mu_i,\Sigma_i), \tag{91}$$

where $a=(a_1,\ldots,a_{\boldsymbol{K}})$ are the normalized weights of the Gaussian distributions, i.e., $\sum_{i=1}^{\boldsymbol{K}} a_i = 1$ and $a_i \ge 0,\ i=1,\ldots,\boldsymbol{K}$. Now notice that the upper bound in (91) coincides with the average cluster function induced by (90) if we assume that in (54) $\boldsymbol{C}=0$, i.e., $\Gamma(t)\equiv a\ \forall t$. However, the associated optimal affiliation function,

$$a_i^* = \begin{cases} 1 & \text{if } i = \operatorname{argmin}_j\big\{-\sum_t \log\rho_G(x_t;\mu_j,\Sigma_j)\big\},\\ 0 & \text{otherwise,}\end{cases} \tag{92}$$

is deterministic, implying that the optimal substitute model (Gaussian mixture model) consists only of one locally stationary model (Gaussian distribution) *independent* of the number $\boldsymbol{K}$ of assumed clusters.

In contrast, the update formula for the weights $a_1, \ldots, a_{\boldsymbol{K}}$ in the classical GMM framework (see, e.g., [61]),

$$a_i^{(s+1)} = \frac{1}{T+1} \sum_{t=0}^{T} \frac{q(i,t)}{\sum_{j=1}^{\boldsymbol{K}} q(j,t)} \quad \text{with } q(i,t) = a^{(s)} \log \rho_G(x_t; \mu_i^{(s)}, \Sigma_i^{(s)}), \quad (93)$$

significantly differs from (92) and, generally, does not lead to a degenerated (deterministic) cluster affiliation as in the FEM-BV approach presented above.

This observation allows the conclusion that the upper bound derived in the GMM framework is sharper than the corresponding average cluster function, e.g., in the right-hand side of (91). However, the assumption of stationary weights ($\boldsymbol{C} = 0$) deployed in the GMM framework is very restrictive and it is not fulfilled in many applications.

3. Model selection

The outcome of the FEM-BV methodology crucially depends on the specific choice of the number of clusters $\boldsymbol{K}$ and the persistence threshold $\boldsymbol{C}$ as the choice expresses a certain *a priori* knowledge on the nature of the data under consideration. In fact, the identification of an optimal or best model among a set of possible models is an important part of the clustering procedure itself. In this section we briefly discuss several approaches that have been proposed in the context of the FEM-BV methodology for the selection of the optimal parameters. Furthermore, we present an extension of a recently introduced *information-theoretical* framework that allows the *simultaneous* identification of the optimal parameters $\boldsymbol{K}$ and $\boldsymbol{C}$.

The characterization of an optimal model in terms of its parameters $\boldsymbol{K}$ and $\boldsymbol{C}$ on the basis of the average cluster function, $\boldsymbol{L}(\boldsymbol{K}, \boldsymbol{C})$, is hampered by the following fact: if the number of clusters and the number of allowed transitions between them is increased then the corresponding a priori knowledge is less restrictive and, therefore, the value of the $\boldsymbol{L}(\boldsymbol{K}, \boldsymbol{C})$ decreases. Particularly, $\boldsymbol{L}(\boldsymbol{K}, \boldsymbol{C})$ attains its minimum in the limit $\boldsymbol{K} = N, \boldsymbol{C} = \infty$, which would imply that the corresponding model is optimal in the sense that it explains the data best. As explained in Section 2.d, however, the resulting model is meaningless due to the over-fitting and does not reveal any insights in the underlying data. Therefore, a criterion for selecting the *optimal* parameters should take both into account: how well the data is explained and the total number of involved parameters such as the number of clusters, the actual number of transitions between the clusters and the number of model parameters in each cluster.

Several approaches have been proposed to tackle the problem of selecting an optimal model within the context of the FEM-BV methodology. For instance, the approach in [63] is based on the following observation. The increase of the number

of clusters leads to an increase of uncertainty of the estimated model parameters for each cluster as less data is assigned. Consequently, if one starts with a large number of clusters, then this number can be reduced by combining the clusters whose parameters have a nonempty intersection of their confidence intervals as those clusters are statistically not distinguishable. The procedure is terminated if all clusters are statistically distinguishable.

To choose the optimal persistence threshold $\boldsymbol{C}$, techniques such as the L-Curve method [50] can be applied. The idea is to analyze the graph of the average clustering functional as a function of the persistence threshold $\boldsymbol{C}$. The optimal $\boldsymbol{C}^*$ is then characterized by the point of maximum curvature of the graph.

Recently in [33], an information theoretical framework has been introduced for the *simultaneous* identification of the optimal parameters $\boldsymbol{K}^*$ and $\boldsymbol{C}^*$. It is motivated by the principle of *Occam's razor*: the best or optimal model among a set of possible models is the one that exhibits maximal model quality (goodness of fit) while its number of free parameters is minimal. The most prominent information measure embodying that principle is the AIC (Akaike information criterion, introduced in [1]), which, formally, is given by

$$\mathrm{AIC}(M) = -2\ln \mathscr{L}(M) + 2|M|, \tag{94}$$

where $\mathscr{L}(M)$ denotes the *likelihood* of the model M and $|M|$ is the total number of the model's free parameter. The optimal model M^* is then characterized by the one that minimizes the criterion.

The AIC depends on the likelihood $\mathscr{L}(M)$ of the model as a measurement of the model quality. Therefore, the criterion can not be generally applied in the FEM-BV methodology because it is based on the more general notion of a *model distance function*.

If the model distance function, however, is induced by, e.g., a discrete probability distribution (cf. (34) in Section 2.c.iii) then as justified by Theorem 2.2 (see Section 2.h) the likelihood $\mathscr{L}(M)$ reduces to the likelihood given in (82). Analogously, the reasoning carries over to a model distance function defined in terms of a PDF (cf. (90) in Section 2.h) and to a model function preserving probability such as the Markov regression model introduced in Section 2.c.iv.

It remains to consider the case, e.g., FEM-BV-k-means, if neither the model function nor the model distance function allows a probabilistic interpretation. Fortunately, the gap can be bridged by realizing that the *distribution of the scalar time series of model distances with respect to a fixed cluster i* reflects how well the corresponding local model explains the data. The key idea now is to employ these distribution in order to define a likelihood of a scalar process and, eventually, to arrive at a modified information criterion for detecting the optimal model in the FEM-BV approach.

Let $\mathrm{supp}(\gamma_i) = \{t : \gamma_i(t) > 0\}$ denote the support of $\gamma_i(t)$ and suppose for a moment that the model distances in the cluster $i = 1, \ldots, \boldsymbol{K}$ are each distributed according to a parametric (conditional) probability density function (PDF) $\rho_i(\cdot\,; \Lambda_i)$, i.e.,

$$\mathbb{P}[g(x_t, \theta_i) \in \mathrm{d}x] = \rho_i(g(x_t, \theta_i);\, \Lambda_i)\,\mathrm{d}x, \quad i = 1, \ldots, \boldsymbol{K},\ \forall t \in \mathrm{supp}(\gamma_i). \tag{95}$$

Under the (restrictive) assumption of independence, we can define a likelihood function $\mathscr{L}(\boldsymbol{K}, \boldsymbol{C})$ by

$$\mathscr{L}(\boldsymbol{K}, \boldsymbol{C}) \stackrel{\mathrm{def}}{=} \prod_t \Bigg(\sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) \rho_i(g(x_t, \theta_i);\, \Lambda_i) \Bigg) \tag{96}$$

and, following the arguments from the original proof by Akaike [1], we arrive at the modified information criterion

$$mAIC(\boldsymbol{K}, \boldsymbol{C}) = -2\ln(\mathscr{L}(\boldsymbol{K}, \boldsymbol{C})) + 2|M(\boldsymbol{K}, \boldsymbol{C})|. \tag{97}$$

The total number of the model's free parameters, $|M(\boldsymbol{K}, \boldsymbol{C})|$, consists of three contributions; the total number of local stationary parameters, i.e., $|\Theta| = |\theta_1| + \cdots + |\theta_{\boldsymbol{K}}|$, the total number of parameters needed for describing the conditional PDFs, i.e., $|\Lambda| = |\Lambda_1| + \cdots + |\Lambda_{\boldsymbol{K}}|$ and, finally, the total number of parameters needed to represent the affiliation function Γ. To determine $|\Gamma|$, please recall that Γ is piecewise constant on a FEM-partition $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_{N-1} < \tau_N = T$ dividing the interval $[0, T]$ into N bins (Section 2.f). Hence, we conclude

$$|\Gamma| = \boldsymbol{K}N. \tag{98}$$

For instant, the total number of parameters in the FEM-BV-k-means model is (Section 2.c.i)

$$|M_{k\text{-means}}(\boldsymbol{K}, \boldsymbol{C})| = \boldsymbol{K}d + \boldsymbol{K}N + |\Lambda|. \tag{99}$$

It remains to explain how to characterize the set of parametric PDFs, $\{\rho_i(\cdot\,; \Lambda_i)\}$, capturing the respective distribution of the cluster's model distances appropriately. One option is to *assume* that all distributions during the course of optimization belong to a certain but fixed class of parametric PDFs, e.g., the class of Gaussians. The parameters Λ_i are then efficiently calculated via the maximum likelihood approach. However, our numerical experiments showed that the assumption of a fixed class of parametric PDFs is too restrictive and may lead to wrong optimal models.

To motivate the approach presented here, note that we actually do not know anything about the parametric representations of the distributions. What we can empirically compute, however, are statistical properties such as the expectation, the variance and, more generally, the first k noncentralized moments. The key idea now

is to choose the *most unbiased distribution* in each case, among those exhibiting the empirical observed statistical properties. According to [43; 44; 55] the most unbiased distribution is the one which admits the *most uncertainty measured in terms of entropy.*

Let η_j, $j = 0, \ldots, k$ be empirical estimates of the first $k+1$ noncentralized moments of a distribution with $\eta_0 = 1$. The associated *maximum entropy distribution* is characterized by a constrained variational problem

$$\mathcal{H}(\rho) \stackrel{\text{def}}{=} -\int \rho(x) \ln \rho(x)\, \mathrm{d}x \to \max_{\rho(x) \in L^2(\mathbb{R})} \tag{100}$$

subject to

$$\eta_j = \int x^j \rho(x) \mathrm{d}x, \quad j = 0, \ldots, k, \tag{101}$$

where $\mathcal{H}(\rho)$ is the entropy of the PDF ρ.

Applying the calculus of variation yields the formal (unique) solution

$$\rho^*(x) = \exp \sum_{j=0}^{k} \lambda_j x^j = \operatorname*{argmax}_{v(x) \in L^2(\mathbb{R})} \mathcal{H}(\rho), \tag{102}$$

where the Lagrange multipliers $\lambda_0, \ldots, \lambda_k$ enforce the constraints in (101). For instant, if $k = 2$ then ρ^* is basically given by a Gaussian distribution having the prescribed moments. Unfortunately, for $k > 2$ no closed expression for ρ^* exists so that the Lagrange multipliers have to be computed numerically via, e.g., the Newton method. For details on solving the problem (100)–(101) numerically see, e.g., [76]. Moreover, for an overview on maximum entropy distributions associated with constraints other than in (101) we refer to, e.g., [46; 55].

The maximum entropy ansatz finally allows us to characterize the parametric representations of the distributions of the respective (scalar) cluster's model distances

$$\{g(x_t, \theta_i)\}, \quad t = 0, 1, \ldots, T, \ i = 1, \ldots, \boldsymbol{K} \tag{103}$$

as

$$\rho_i(x, \lambda_0^{(i)}, \ldots, \lambda_k^{(i)}) = \exp \sum_{j=0}^{k} \lambda_j^{(i)} x^j \tag{104}$$

subject to

$$\int x^j \rho_i(x, \lambda_0^{(i)}, \ldots, \lambda_k^{(i)}) \mathrm{d}x = Z_i^{-1} \sum_{t \in \mathrm{supp}(\gamma_i)} (g(x_t, \theta_i))^j \quad j = 0, \ldots, k, \tag{105}$$

with $Z_i = |\mathrm{supp}(\gamma_i)|$. Inserting (104) in (94) we end up with the *modified* AIC, denoted by $mAIC(\boldsymbol{K}, \boldsymbol{C})$, for selecting the optimal model within the FEM-BV

methodology. Notice that we only require in (104) and (105) the scalar "observables" $g(\cdot, \theta_i)$ to be i.i.d.[2] Furthermore, the optimal number (order) k of moments needed to approximate the underlying distribution can again be determined by employing the AIC.

We end this section by discussing a conceptual weakness of the presented model selection approach. Despite its successful application and the numerical evidence indicating its usefulness (see Section 5 below), the approach theoretically suffers from the fact that the estimation of the ME-distributions is invariant under translation, i.e., the ME-distributions estimated from, e.g., the scalar time series $(g(x_t, \theta^*)), t \geq 0$ and $(g(x_t, \theta^*) + a),\ a > 0, t \geq 0$ would be indistinguishable from the view point of likelihood. Consequently, they would equally contribute to the modified AIC although the former distribution is closer to the lower bound, (say zero), and, hence, the associated underlying model should be the preferred one. From the practical point of view, such scenarios are very unlikely to happen since the model distance function $g(x_t, \theta)$ is minimized during the subspace-procedure. In fact, the occurrence of such a scenario would indicate that the underlying model function $f(\cdot, \theta(t))$ does not properly capture the dynamic of the time series under consideration.

Generally spoken, the model selection approach theoretically suffers from not explicitly incorporating the lower boundedness of the model distance function $g(\cdot)$. Bridging that gap is subject to ongoing research and will be discussed in a forthcoming manuscript.

4. Self-containing predictive models

In the previous section, we presented for the FEM-BV approach a tailored strategy to identify an optimal stochastic model in terms of the optimal number of clusters $\boldsymbol{K}^*$ and the optimal persistence $\boldsymbol{C}^*$. Furthermore, we elaborated in Section 2.g under which conditions the optimal model parameters and the optimal cluster affiliations lead to a time-dependent mixture model for fitting the data best within the trained time interval. In this section we present a *prediction strategy* allowing us to predict the dynamics beyond the trained time interval.

Let $\Gamma^*_{[m,T]}$ and $\theta^*_1, \ldots, \theta^*_{\boldsymbol{K}^*}$ be the parameters of the optimal model associated with a model function $f(\cdot, \cdot)$ on the time interval $[m, T]$. A reasonable fitting (prediction) at $t \in [m, T]$, i.e., within the trained time span, is then given by the

[2] In this context it is important to recall the standard application of information functionals for Bayesian time series analysis methods (such as GMMs and HMMs) [21; 49] relies on a very restrictive additional assumption, namely that the analyzed data x_t are produced by a known parametric multivariate distribution.

average mixture model (cf. (81) in Section 2.g)

$$\hat{x}_t = \mathbb{E}\left[f\left(x_t, x_{t-1}, \ldots, x_{t-m}, \sum_{i=1}^{\boldsymbol{K}^*} \gamma^*_{i,[m,T]}(t)\theta^*_i \right) \right]. \tag{106}$$

Now it is important to realize that the average mixture model is confined on the interval $[m, T]$ because it explicitly depends on the time-dependent affiliation function $\Gamma^*_{[m,T]}$ being only well defined on $[m, T]$. However, if we could predict the affiliation function $\hat{\Gamma}_{[m,T+d]}(t)$ for $t = T+1, \ldots, T+d,\ d > 0$ then (106) could readily be extended for predicting $\hat{x}_{T+1}, \ldots, \hat{x}_{T+d}$ by the following recursive scheme

$$\hat{x}_{T+r} = \mathbb{E}\left[f\left(\hat{x}_{T+r}, \ldots, \hat{x}_{T+r-m}, \sum_{i=1}^{\boldsymbol{K}^*} \hat{\gamma}_{i,[m,T+d]}(T+r)\theta^*_i \right) \right] \quad r = 1, \ldots, d \tag{107}$$

with $\hat{x}_s = x_s$ and $\hat{\Gamma}_{[m,T+d]}(s) = \Gamma^*_{[m,T]}(s)$ if $s \leq T$.

A self-contained strategy for predicting $\hat{\Gamma}_{[m,T+d]}$ has been recently proposed in [34]. It is based on two simple but fundamental observations. Firstly, $\Gamma^*_{[m,T]}(t),\ t = m, \ldots, T$ itself can be viewed as a time series of discrete probability distributions due to the imposed convexity conditions in (13). Secondly, under the assumption that the *distributions* $\Gamma^*_{[m,T]}(t),\ t = m, \ldots, T$ are associated with a (hidden) time-homogeneous and stationary Markov process, a model for the dynamics of the cluster affiliations is readily given by ($\Gamma \equiv \Gamma^*_{[m,T]}$)

$$\Gamma^\dagger(t+1) = \Gamma^\dagger(t)P, \tag{108}$$

where $P \in \mathbb{R}^{\boldsymbol{K}^* \times \boldsymbol{K}^*}$ is a stochastic matrix, i.e., P is elementwise nonnegative and the entries of a row sum up to 1.

Particularly, the dynamics in (108) allows the recursive prediction of $\hat{\Gamma}_{[m,T+d]}$, e.g.,

$$\hat{\Gamma}^\dagger(T+1) = \hat{\Gamma}^\dagger(T)P \tag{109}$$

with $\hat{\Gamma}(T) = \Gamma^*_{[m,T]}(T)$ and, finally, in combination with (107) leads to a self-contained prediction scheme for the dynamics of the data under consideration.

This leaves us with the question how to estimate the stochastic matrix P from the time series of affiliations. Of course, in general we can not expect that a matrix P exists such that (108) exactly holds true. However, the FEM-BV methodology, in particular the approach presented in the Section 2.c.iv, provides an elegant way to deal with that situation by solving the following variational problem (cf. (42)):

$$\sum_{t=0}^{T-1} \left\| \Gamma^\dagger(t+1) - \Gamma^\dagger(t)P \right\|_2^2 \to \min_P, \tag{110}$$

subject to P being a stochastic matrix and $\Gamma \equiv \Gamma^*_{[m,T]}$. In order to study the influences of external factors $u(t) \in \mathbb{R}^k$, we additionally assume that the matrix $P = P(u(t))$ can be decomposed as (cf. (44))

$$P(u(t)) = P_0 + \sum_{l=1}^{k} u_l(t) P_l, \tag{111}$$

where the involved matrices satisfy the constraints (46)–(48). The final minimization problem (110)–(111) with respect to the parameters $P_0, \ldots, P_k \in \mathbb{R}^{\boldsymbol{K}^* \times \boldsymbol{K}^*}$ and subject to the constraints (46)–(48) can be cast in a constrained quadratic program. For details see the Appendix.

Next, suppose that an observation for x_{T+1} is available. What is the optimal prediction for $\hat{x}_{T+2}$ *conditioned* on the additional observation x_{T+1}? As motivated in [34], instead of reanalyzing the updated time series $(x_0, \ldots, x_{T+1})$ via the FEM-BV approach and reapplying the prediction scheme described above, it is sufficient to determine the optimal affiliation vector $\Gamma^*_{[m,T+1]}(T+1)$ simply by

$$\gamma_i^*(T+1) = \begin{cases} 1 & \text{if } i = \operatorname{argmin}_j \left\{ g(x_{T+1}, \ldots, x_{T+1-m}, \theta_j^*) \right\}, \\ 0 & \text{otherwise,} \end{cases} \tag{112}$$

which by virtue of (108) and (107) yields the prediction $\hat{x}_{T+2}$. The generalization of the conditional prediction in the presence of more than one new observation, say $(x_{T+1}, \ldots, x_{T+t'})$, is straightforward. The resulting scheme is $(r = 1, \ldots, d)$

$$\begin{aligned} (x_{T+t'}, \Theta^*) &\overset{\text{via (112)}}{\longrightarrow} \Gamma^*_{[m,T+t']}(T+t') \\ &\overset{\text{via (108)}}{\longrightarrow} \hat{\Gamma}_{[m,T+t'+d]}(T+t'+r) \overset{\text{via (107)}}{\longrightarrow} \hat{x}_{T+t'+r}. \end{aligned} \tag{113}$$

The remainder of this section is devoted to describing numerical strategies to assess the prediction quality of the scheme given above. To this end, we compare $\hat{x}_{T+k}$ with standard prediction approaches such as the "zero" prediction model frequently used in, e.g., the meteorological literature. Formally, it reduces to

$$\hat{x}^0_{T+d} \equiv x_T. \tag{114}$$

Furthermore, as frequently pointed out in this manuscript, stationarity is a widely used and well accepted assumption in time series analysis. Thus, it is reasonable to compare $\hat{x}_{T+d}$ with the prediction $\hat{x}^1_{T+d}$ resulting from an optimal *stationary* substitute model, i.e. (analogously to (107))

$$\hat{x}^1_{T+d} = \mathbb{E}\left[f\left(\hat{x}^1_{T+d}, \ldots, \hat{x}^1_{T+d-m}, \theta^* \right) \right], \tag{115}$$

where θ^* is derived[3] from the time series under consideration.

[3]Numerically, this simply amounts to fix $\boldsymbol{K} = 1$ in the course of the FEM-BV approach.

The *average relative prediction error* of the d-step prediction scheme for a prediction horizon $[T+1, T+T']$ is then measured by

$$\bar{e}_d(T') \stackrel{\text{def}}{=} \frac{1}{T'-d+1} \sum_{t'=T}^{T'-d} \frac{\left\| x_{t'+d} - \hat{x}_{t'+d} \right\|}{\| x_{t'+d} \|}, \tag{116}$$

where $\|\cdot\|$ denotes a desired norm. That error is compared with the average relative error $\bar{e}_d{}^0(T')$ associated with the zero-prediction scheme and $\bar{e}_d{}^1(T')$ resulting from predicting via the stationary substitute model. See Section 5.e for a numerical example illustrating the described prediction schemes. Another possibility for measuring the prediction error is given by the information-theoretical approaches to model error assessment developed at the working group of A. Majda (NYU); we refer the interested reader to, e.g., [55; 24] for more details on this matter.

5. Numerical examples

In this section we illustrate the presented FEM-BV methodology on various examples. In the first and second example we demonstrate the general feasibility of the proposed method and discuss its properties on a simple model with known properties. In the third example, a modified version of the FEM-BV-k-means for periodic angular data is developed and applied to analyze the conformational dynamics of a small biomolecule. The fourth example deals with a problem in computational biology and shows that the FEM-BV framework adapted for discrete data allows us to analyze gene-sequences under minimal a priori assumptions. The analysis of financial data is presented in the last example in which we also discuss the usefulness of the self-contained prediction scheme presented in Section 4.

5.a. ***Toy model system I: FEM-BV-k-means.*** The k-means approach is a widely used algorithm to cluster stationary data on the basis of geometric properties, i.e., the Euclidean distance to geometric centroids. However, even for low dimensional examples k-means fails to identify the "right" clusters. In the first numerical experiment we show for such a counter example that the additional information of the temporal (persistent) ordering of the data is sufficient to separate geometric clusters via of FEM-BV-k-means.

To this end we consider a time series of two dimensional data $x(t) = (x_1(t), x_2(t))$ generated via a mixture model consisting of a time dependent convex combination of three (stationary) normal distributions,

$$x_t \sim \sum_{i=1}^{3} \gamma_i(t) \mathcal{N}(\mu_i, \Sigma_i) \quad t = 1, \ldots, 6000, \tag{117}$$

where the weights (cluster affiliations) $\Gamma(t) = (\gamma_1(t), \gamma_2(t), \gamma_3(t))$ are deterministic

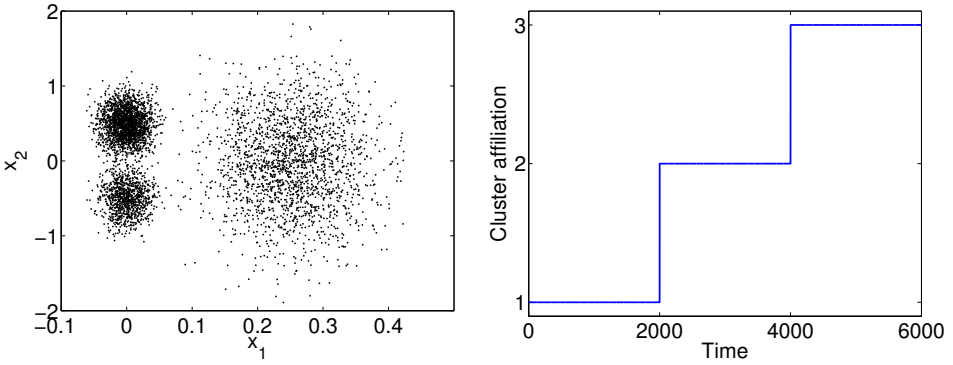


Figure 2. Toy model I. Left: scatter plot of the a time series generated via (117) where the parameters were chosen such that the data exhibits three geometric clusters. Right: the graph of the cluster affiliations used as a persistent hidden process in parameter space for the generation of the time series depicted in the left panel.

and prescribed. Particularly, $\Gamma(t)$ was chosen such that the (hidden) affiliation process jumps only once from cluster one to cluster two and finally to cluster three, i.e., $\|\gamma_1\|_{BV} = \|\gamma_3\|_{BV} = 1$ and $\|\gamma_2\|_{BV} = 2$. For an illustration of $\Gamma(t)$ see the right panel of Figure 2. As one can see in the scatter plot given in the left panel of Figure 2, the means and covariance matrices (μ_i, Σ_i), $i = 1, 2, 3$ were chosen such that a sufficiently long sample (here $T = 6000$) exhibits three geometrically nonoverlapping clusters. However, the k-means algorithm for $k = 3$ failed to identify these clusters as illustrated in the left panel of Figure 3. Notice that the misclassification of the data points is basically due to the different scales of the x_1 and x_2 components of the data.

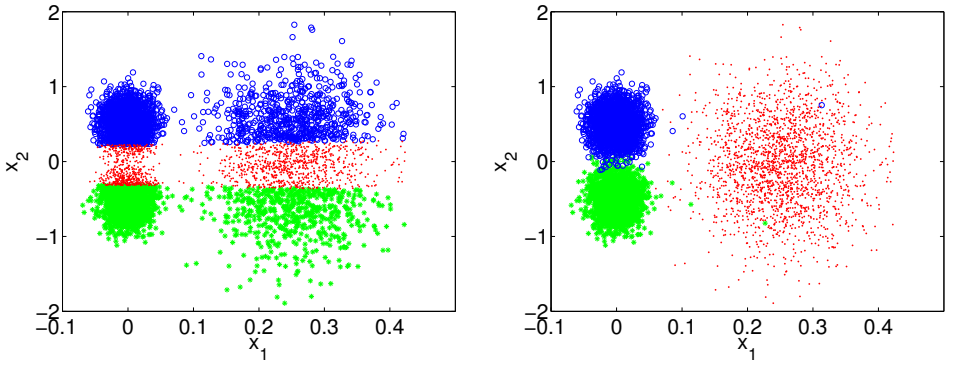


Figure 3. Toy model I. Cluster affiliations of the data points resulting from the classical k-means algorithm (left) and from the FEM-k-means method (right). Up to a few misclassifications, the latter method led to the right assignment of the data points to the original clusters, whereas the former one totally messed up.

Next, we analyzed the time series with the FEM-BV approach which results from the simple model in (6) and the model distance function in (9). Recall that for $\boldsymbol{C} = \infty$ the average cluster functional admits an analytic solution for Γ and for the cluster parameter Θ which both coincide with the respective update formulas in the k-means algorithm (Section 2.a). Now the question is whether the persistence of the prescribed cluster affiliations is sufficient to identify the three cluster while using the same distance function as in the standard k-means approach?

To this end, we repeatedly launched the FEM-BV-k-means subspace algorithm (cf. Section 2.b and (72)–(75)) for all combinations of

$$\boldsymbol{K} = [2, 3, 4] \times \boldsymbol{C} = [2, 4, \ldots, 12],$$

each time with a randomly drawn initial Γ, until the global minimizer of the average cluster functional was found. For the respective optimal models we then computed the modified AIC values via the Maximum-Entropy approach presented in Section 3. For fixed $\boldsymbol{K}$ the graphs of $mAIC(\boldsymbol{K}, \boldsymbol{C})$ as a function of $\boldsymbol{C}$ are given in the left panel of Figure 4. The overall minimum is attained in $\boldsymbol{K}^* = 3$, $\boldsymbol{C}^* = 2$ which are exactly the parameters of the original data. In the right panel of Figure 4, we exemplarily illustrate the histogram of the residuals (6) of the right geometrical cluster together with the graph of the fitted ME PDF (102) of order 3 which was used to compute the modified AIC values. Finally, the correct (up to a few isolated misfits) assignments of data points to the clusters based on the affiliation vector $\Gamma(t)$ is given in the right panel of Figure 3.

This simple but instructive example demonstrates that neglecting temporal persistence in data may lead to misleading results even for toy examples. In contrast, besides yielding the correct partition of the data, the FEM-k-means-method

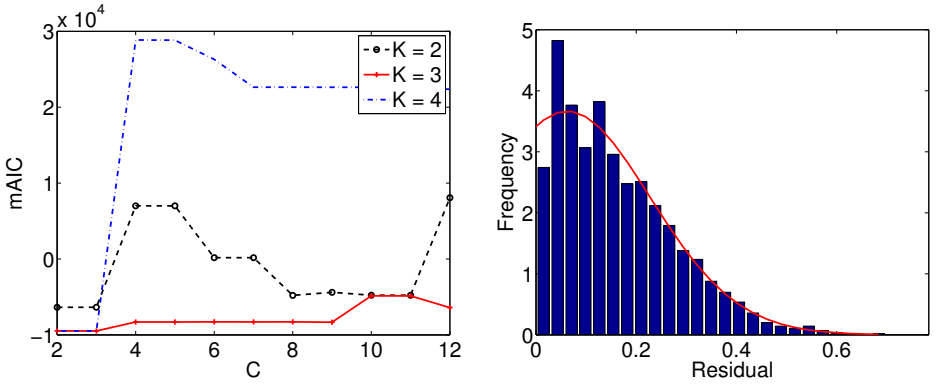


Figure 4. Toy model I. Left: graphs of the (modified) AIC values (97) for fixed $\boldsymbol{K}$ as a function of $\boldsymbol{C}$. Right: the histogram of the residuals (6) of the right geometrical cluster together with the graph of the fitted ME PDF (102) of order 3 (red line).

combined with the model selection approach allowed us to reidentify the correct parameters $\boldsymbol{K} = 3, \boldsymbol{C} = 2$.

5.b. ***Toy model system II: FEM-BV-PCA.*** In the first example, the geometrical clustering of the time series basically relied on the separability via centroids, i.e., mean values. In the second example we demonstrate that even geometric cluster with comparable means can be reidentified via the FEM-BV approach by additionally incorporating spectral properties of covariances, i.e., principal components.

To this end, we consider time series of two dimensional data $x(t) \in \mathbb{R}^2$ of length $T = 10000$ generated via

$$x_t \sim \gamma_1(t)\mathcal{N}_2(0, \Sigma_1) + \gamma_2(t)\mathcal{N}_2(0, \Sigma_2). \tag{118}$$

The prescribed weights (cluster affiliations) $\Gamma(t) = (\gamma_1(t), 1 - \gamma_1(t))$ are deterministic. For an illustration of $\gamma_1(t)$ see the right panel in Figure 5. The covariance matrices Σ_1 and Σ_2 are chosen as

$$\Sigma_1 = \begin{bmatrix} 4 & 0 \\ 0 & 0.25 \end{bmatrix}, \Sigma_2(\rho) = \begin{bmatrix} \cos\rho & \sin\rho \\ -\sin\rho & \cos\rho \end{bmatrix} \Sigma_1 \begin{bmatrix} \cos\rho & -\sin\rho \\ \sin\rho & \cos\rho \end{bmatrix}, \tag{119}$$

where Σ_2 results from rotating Σ_1 by an angle $\rho = 15$ degrees. The scatter plot of the time series generated via (118) is depicted in the left panel of Figure 5. As one can see, the two clusters are almost identical and, by construction, are centered around $(0, 0)$. Therefore, any k-means clustering approach would fail to recover the original temporal affiliation. The only chance to identify the (hidden) cluster though is to cluster with respect to the eigenvectors of the (hidden) covariance

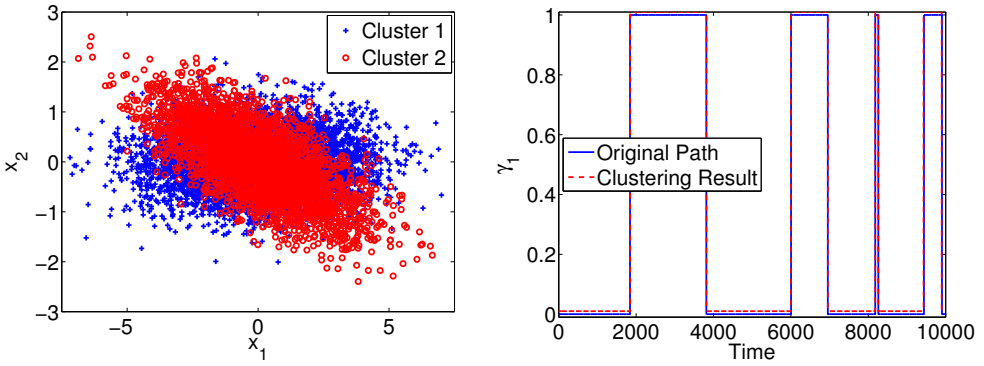


Figure 5. Toy model II. Left: scatter-plot of a time series generated via the mixture model in (118) consisting of a time dependent convex combination of two (stationary) normal distributions with mean zero and covariance matrices given in (119) and a rotation angle $\rho = 15$ degrees. Right: the prescribed affiliation function $\gamma_1(t)$ (solid line) completely coincides with one obtained from the FEM-BV-PCA-analysis (red dashed line).

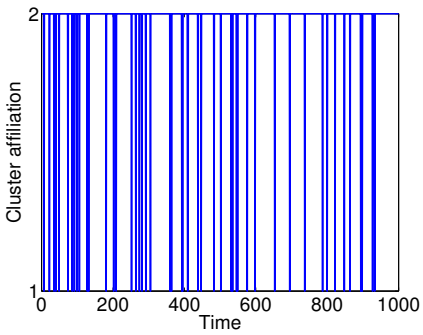


Figure 6. Toy model II: part of the Viterbi path ($1 \leq t \leq 1000$) obtained from fitting a two-dimensional stationary mixture model of two Gaussian distributions (via the GMM-method) on the data shown in Figure 5.

matrices. But this is exactly the idea of the FEM-BV-PCA approach which will be used here.

Before we present the results of the FEM-BV-PCA approach, we first apply the GMM-method which is a classical and widely accepted method for unsupervised clustering. We fitted (trained) a two-dimensional stationary mixture model of two Gaussian distributions on the data via the Expectation-Maximization algorithm [12]. Since Gaussians are involved in the time series generation, it is reasonable to expect the GMM-method to be able to reidentify the parameters of the hidden distributions. However, the estimated covariance matrices $\tilde{\Sigma}_1$ and $\tilde{\Sigma}_2$ significantly differ from the original ones, indicated by, e.g., $\|\Sigma_2 - \tilde{\Sigma}_2\| = 5.2040$.

The associated Viterbi path (partially depicted in Figure 6) reveals the reason for the failure; it is highly oscillatory rather than being persistent. Consequently, the majority of data points are incorrectly affiliated with regard to the original clusters which, ultimately, leads to the incorrect estimation of the covariance matrices. The irregularity of the Viterbi path, in turn, is a direct consequence of the strong stationary assumption underlying the GMM-method, i.e., time-independent distribution parameters and time-independent affiliation weights.

In contrast, as will be demonstrated in the following, the FEM-BV-PCA-method (see Section 2.c.ii) succeeded as it takes the persistence of the hidden dynamics in the parameter space into account. Analogously to the procedure described in the previous example in Section 5.a, we globally minimized the average cluster functional resulting from the model distance function in (27) via the subspace algorithm for all combinations of $\boldsymbol{K} \in \{1, 2, 3\}$ and $\boldsymbol{C} \in \{2, 4, 6, 8, 10, 14, 20\}$. The minimum of the corresponding modified AIC values is attained for $\boldsymbol{K}^* = 2$ and $\boldsymbol{C}^* = 8$, which are exactly the parameters used for the time series generation. Even more importantly, the numerically obtained affiliation vector is identical with the original one (see right panel of Figure 5).

5.c. ***Conformation analysis of a biomolecule (trialanine): FEM-BV-k-means.*** The biological function of a biomolecule is strongly characterized by its ability to assume almost constant geometrical configurations, referred to as *conformations*. More precisely, a conformation denotes a mean geometrical configuration of a molecule which is almost stable (metastable, persistent), i.e., the molecule's geometry wiggles around that configuration for a long period of time before it rapidly switches to another conformation. For example, it is known that conformations of certain proteins are responsible for severe human diseases [51]. For details on the analysis of the conformational dynamics of molecules we refer the interested reader to, e.g., [69] and the references therein.

It is common to analyze the conformational dynamics of a (bio-)molecule in internal coordinates such as torsion angles rather than to consider the time series of cartesian coordinates of all atomic positions. The reason is that torsion angles are invariant with respect to translation and rotation of the molecule and, more importantly, tremendously reduce the dimensionality of the time series. However, the (nonlinear) projection of the cartesian coordinates on the torsion angle space deflects the original dynamics and can lead to an incomplete picture of the conformational dynamics of the molecule. This is in particular true if only a subset of torsion angles is considered because of, e.g., numerical or statistical reasons. Consequently, conformations which are geometrically distinguishable in the complete torsion angle space might (completely) overlap in the reduced space. Thus, the identification of conformations via geometrical clustering of *incomplete* observations of torsion angles is an *ill-posed* problem.

In the traditional *transfer operator (TO) approach* [65] to conformational dynamics the problem is addressed by assuming that the underlying dynamics in the incomplete torsion angle space is a *reversible, stationary and time-homogeneous Markov process*. Alternatively, we propose to tackle the ill-posedness by regularization of the underlying persistent (metastable) dynamics in the BV sense and to identify conformations via a modified FEM-BV-k-means approach.

To this end, we consider in this example a time series of three torsion angles Φ, Ψ and Ω obtained from a molecular simulation of the trialanine molecule schematically illustrated as a ball-stick representation in the left panel of Figure 7. The simulation was performed in vacuum at constant temperature and pressure such that the resulting time series can be considered stationary for a sufficiently long simulation time T. The details of the simulation procedure can be found in [60]. As one can see in the right panel of Figure 7, the dynamics of the torsion angles exhibits a strong persistence or metastability.

Recalling that the torsion angles are periodic on $[-\pi, \pi]$, the 3d-scatter plot in the left panel of Figure 8 clearly reveals five geometrical clusters indicating five conformations. The projection on the two torsion angles Φ and Ψ, however,

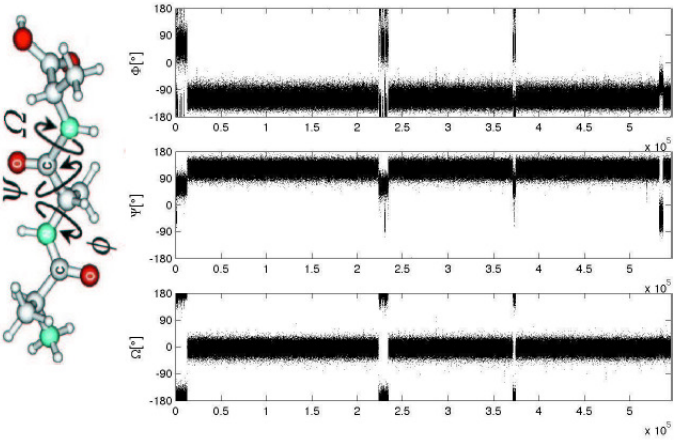


Figure 7. Biomolecule: molecular simulation of the trialanine molecule (left) reveals its conformational dynamics observed in the time series of three torsion angles (right).

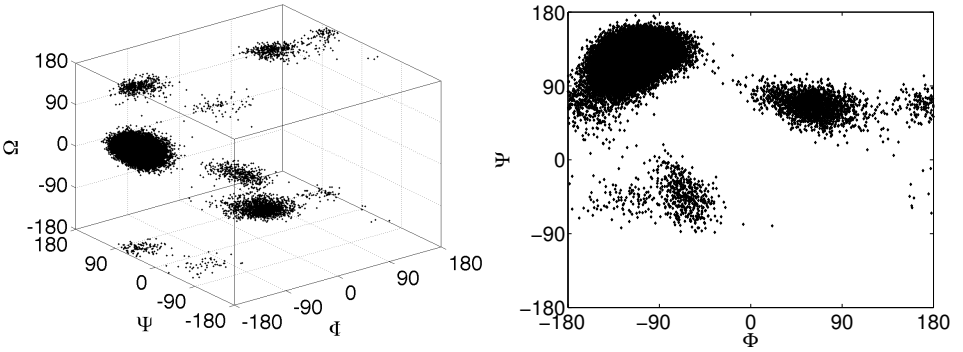


Figure 8. Biomolecule: recalling the periodic nature of torsion angles, the scatter-plot of the full time series (left) reveals five conformational clusters whereas the scatter-plot of the projected time series $(x_t) = (\Phi_t, \Psi_t)$ (right) suggests the existence of only three conformations.

suggests the existence of only three conformations as illustrated in the right panel of Figure 8. Consequently, the five clusters can only be recovered in the projection by additionally capturing the inherent persistence of the dynamics. This will be demonstrated in the remainder of this example.

To understand the following preprocessing steps, we briefly recall the transfer operator approach to conformation dynamics. The basic idea is to represent the dynamics underlying the time series of torsion angles as a *reversible, stationary and time-homogeneous Markov chain* defined on a suitable discretization of the torsion angle space, e.g., by boxes. The spectrum of the associated transition matrix P, then allows the characterization and extraction of the conformations as metastable subsets via, e.g., the robust Perron-cluster cluster analysis (PCCA+) [14]. To be

more precise, let $\lambda_1, \lambda_1, \ldots, \lambda_n$ be the first n dominant eigenvalues of P, i.e.,

$$1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|. \tag{120}$$

If a *spectral gap* exists, i.e., if one can find an index $\boldsymbol{K}$ such that $|\lambda_{\boldsymbol{K}}| \gg |\lambda_{\boldsymbol{K}+1}|$, then one can prove that the discrete state space can be decomposed into $\boldsymbol{K}$ metastable subsets (conformations), say $A_1, \ldots, A_{\boldsymbol{K}}$, based on the corresponding dominant eigenvectors [11; 66; 42; 13]. A measure for the total metastability of the resulting decomposition is then given by

$$\eta(A_1, \ldots, A_{\boldsymbol{K}}) = \sum_{i=1}^{\boldsymbol{K}} P(A_i, A_i), \tag{121}$$

where

$$P(A_i, A_i) = \mathbb{P}[x_1 \in A_i | x_0 \in A_i]$$

is the (time-homogeneous) probability that the dynamics is in A_i after making a transition out of A_i.

Accordingly, to ensure Markovianity while preserving the persistence, the original time series was further subsampled by picking every 10-th time step resulting in a time series $(x_t) = (\Phi_t, \Psi_t)$ of total length $T = 54455$. Then, the 2-dimensional space spanned by the torsion angles Φ and Ψ was discretized into 30×30 equidistantly sized boxes and we ended up with a 372-state Markov chain since only 372 boxes are visited by x_t.

The spectral gap between the third and fourth dominant eigenvalue (see left panel in Figure 9) suggests an optimal decomposition into three clusters showing

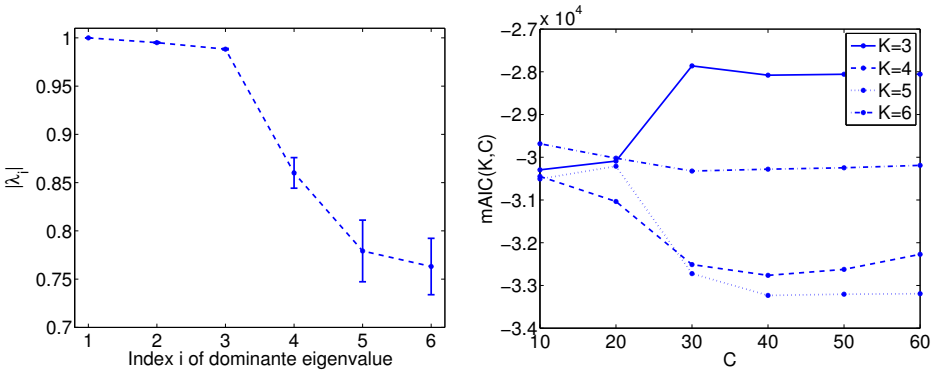


Figure 9. Biomolecule. Left: six dominant eigenvalues of the transition matrix $P \in \mathbb{R}^{372\times372}$ and their confidence intervals, resulting from a 30×30 box discretization of the state space spanned by Φ and Ψ. Right: for fixed $\boldsymbol{K} = 3, 4, 5, 6$ the graphs of the mAIC values as a function of $\boldsymbol{C}$ obtained via the Maximum-Entropy approach with order three. The minimum is attained for $\boldsymbol{K}^* = 5$ and $\boldsymbol{C}^* = 40$.

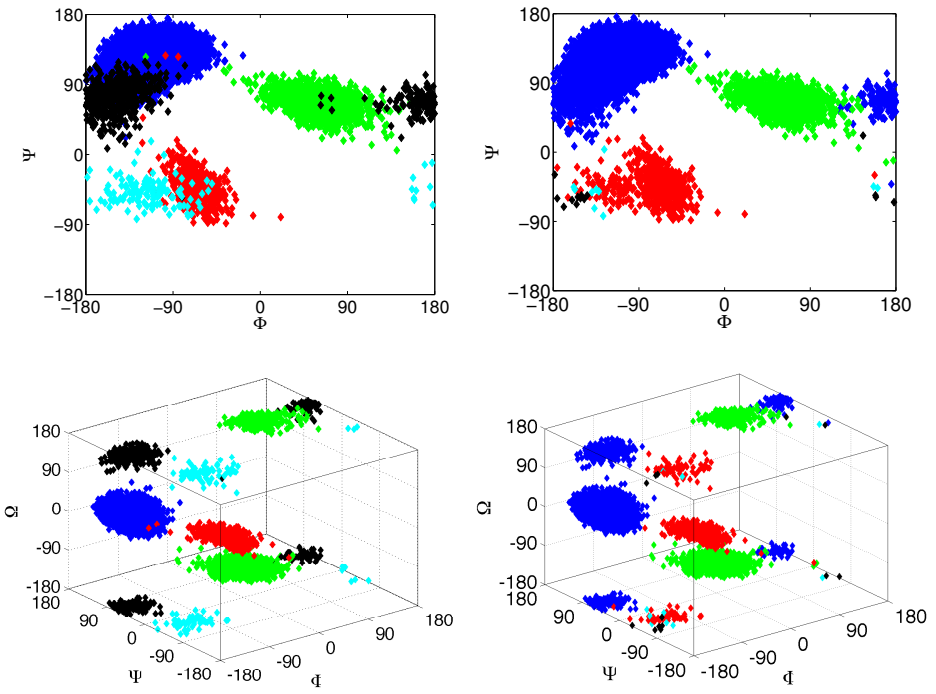


Figure 10. Biomolecule. Decomposition of the time series $(x_t) = (\Phi_t, \Psi_t)$ into five clusters via periodic FEM-BV-k-means (upper left) and the TO approach (upper right). The same decomposition of the time series visualized in a full three-dimensional feature space $(x_t) = (\Phi_t, \Psi_t, \Omega_t)$ reveals the correct identification of the conformations (lower left) by the FEM-BV-k-means method whereas the TO approach is not able to recover them from the incomplete observation in x_t (lower right).

that the TO approach fails to capture the persistence of the dynamics leading to five conformations. Furthermore, as illustrated by the error bars, the high uncertainty[4] of the fifth and sixth dominant eigenvalue indicates that they are statistically indistinguishable and so are the corresponding eigenvectors. Hence, any attempt to decompose the state space into five clusters by additionally considering the fourth and, particularly, the fifth dominant eigenvector would fail to properly separate the conformations. This is confirmed in the right lower panel of Figure 10 and by the fact that the total metastability (121) for the decomposition resulting from the TO approach has the value $\eta_{\mathrm{TO}} = 4.106$, significantly lower than the value $\eta_{\mathrm{FEM}} = 4.900$ resulting from the periodic FEM-BV-k-means method to be presented below.

[4]Based on a 800,000-member transition matrix ensemble generated via a sampling method introduced in [57].

From a more general viewpoint, the high uncertainty in the (less) dominant eigenvalues reflects the ill-posedness of the cluster problem in the presence of incomplete data. Hence, an appropriate regularization is needed such as provided in the variational FEM-BV approach.

As demonstrated in Section 5.a, the simplest way to geometrical clustering while taking persistence into account is the FEM-BV-k-means approach. The model distance function in (9), however, does not capture the *periodic* nature of the data. Fortunately, this can easily be fixed by adopting a distance model function defined on the d-dimensional torus:

$$g(x_t, \Theta_t) = \sum_{j=1}^{d} \left\| \omega([x_t]_j) - \omega([\Theta_t]_j) \right\|_2^2, \quad \text{with } \omega(\alpha) = (\cos\alpha, \sin\alpha) \in \mathbb{R}^2, \tag{122}$$

where $[y]_j$ denotes the j-th component of $y \in \mathbb{R}^d$. A straightforward calculation shows that the average cluster functional associated with (122) attains for given $\Gamma(t)$ a local minimum in $\theta_i^* \in \mathbb{R}^d$, elementwise given by

$$[\theta_i^*]_j = \tan^{-1} \frac{\sum_{t=0}^{T} \gamma_i(t) \sin[x_t]_j}{\sum_{t=0}^{T} \gamma_i(t) \cos[x_t]_j} \quad j = 1, \ldots, d. \tag{123}$$

Via the subspace algorithm, we globally minimized the average cluster functional resulting for all combinations of $\boldsymbol{K} \in \{3, 4, 5, 6\}$ and $\boldsymbol{C} \in \{10, 20, \ldots, 60\}$. The mAIC values are plotted in the right panel of Figure 9. The overall minimum is assumed in $\boldsymbol{K}^* = 5$ and $\boldsymbol{C}^* = 40$ suggesting the existence of five conformations. Indeed, the according decomposition of the full time series (left lower panel of Figure 10) based on the 2-dimensional clustering (left upper panel of Figure 10) shows that the FEM-BV approach succeeded in identifying the conformations most correctly.

In this example we have demonstrated that the FEM-BV-k-means approach adapted for periodic data allows us to identify all of the relevant conformations of a biomolecule based on incomplete torsion angle observations. In particular, we have shown that the combination of BV-regularization with the model selection via the modified AIC does not only yield the correct number but also the correct assignment of the analyzed data to proper conformations. In contrast, although the underlying assumptions necessary for formal applicability of the TO approach (e.g., homogeneity and Markovianity) are formally fulfilled for the analyzed time series, it was demonstrated that the classical transfer operator approach can suffer from the ill-posedness of the clustering problem resulting from the strong overlapping of different conformational states in the reduced representations. The current example

demonstrates that this ill-posedness can result in misleading conformational decompositions in context of the TO approach.

5.d. ***Yeast DNA.*** One of the major challenges in bioinformatics is the identification of genes from biological data. In this example, we approach that problem with the FEM-BV-categorical method derived in Section 2.c.iii and compare the results to classical methods such as the unsupervised HMM.

Gene finding is the identification of coding (*exons* or *genes*) and noncoding (*introns*) regions in nucleic acids (DNA and RNA) based on sequences of codons which specify the amino acid production during the protein synthesis. A codon is a sequence of three nucleotides out of the four possible nucleic bases adenine (A), guanine (G), thymine (T) and cytosine (C). Thus, a single codon can code for a maximum of 64 amino acids.

Traditional approaches to gene finding are based on *supervised machine learning* methods such as supervised HMMs [49], and rely on extensive previous training and a high amount of a priori biological knowledge. Particularly, it is assumed that the hidden process switches exactly between two states, coding and noncoding regions, and that it is a stationary Markov process.

In contrast to the supervised methods, we propose the FEM-BV approach based on the categorical model introduced in Section 2.c.iii as an *unsupervised* approach. We exemplify the usefulness of the method by clustering a sequence c_t of $T = 10'000$ codons resulting from the first $30'000$ nucleotides of the first chromosome of *Saccharomyces cerevisiae*, the ordinary yeast. The data is publicly available at [59]. Notice that in the variational approach the assumption of *persistence* corresponds to the biological assumption that coding and noncoding regions are each *connected*.

After identifying each codon c_t with a discrete state $s_t \in \mathcal{S} = \{1, \ldots, 64\}$ we globally minimized the average cluster functional in (35) (resulting from the model distance function in (34)) for all combinations of $\boldsymbol{K} \in \{1, \ldots, 3\}$ and $\boldsymbol{C} \in \{4, \ldots, 10, 15, 20, 30, 40, 50\}$. Unlike to the previous examples where we applied the Maximum-Entropy approach, here we computed the likelihood function $\mathcal{L}(\boldsymbol{K}, \boldsymbol{C})$, involved in the modified AIC value (97), by exploiting that the stationary cluster parameters $\theta_1, \ldots, \theta_{\boldsymbol{K}}$ are probability distributions. Consequently, $\mathcal{L}(\boldsymbol{K}, \boldsymbol{C})$ takes the form,

$$\mathcal{L}(\boldsymbol{K}, \boldsymbol{C}) = \prod_{t=1}^{T} \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t)\theta_i(s_t). \tag{124}$$

As one can see in Figure 11, the optimal substitute model is attained for $\boldsymbol{K}^* = 2$ and $\boldsymbol{C}^* = 6$. The interpretation of the two clusters as a *coding and a noncoding* model is substantiated by comparing the associated affiliation function $\gamma_1(t)$ with known positions of the genes in this part of the DNA sequence. As one can see

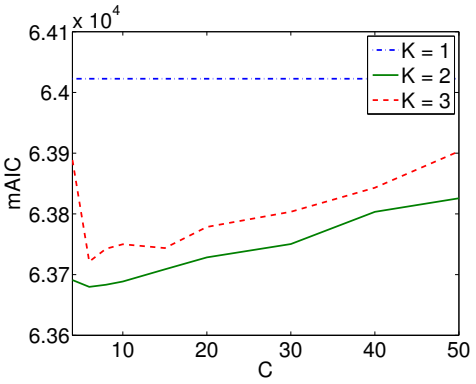


Figure 11. Yeast DNA. The minimal mAIC value is assumed for $\boldsymbol{K}^* = 2$ and $\boldsymbol{C}^* = 6$ which, particularly, is consistent with the biological fact that codons can be divided into coding and noncoding regions.

in the left panel of Figure 12, the affiliation path $\gamma_1(t)$ of the first cluster separates mostly correctly between genes and noncoding regions. Only the gene *SEO1* is not identified which is in contrast to its graphical appearance and length in the left panel of Figure 12. This conflict, however, can be resolved by the experimental fact that this particular region encodes a protein but it does not exhibit a persistent sequence of coding codons because it is highly fragmented, for details see [59]. This violates the persistence assumption inherent to the FEM-BV methodology.

From considering the highly oscillatory Viterbi path of an unsupervised two-state HMM fitting (see right panel of Figure 12) one sees that the assumption of stationarity impedes the traditional approaches to identify genes correctly unless a large amount of biological knowledge is incorporated via supervised learning

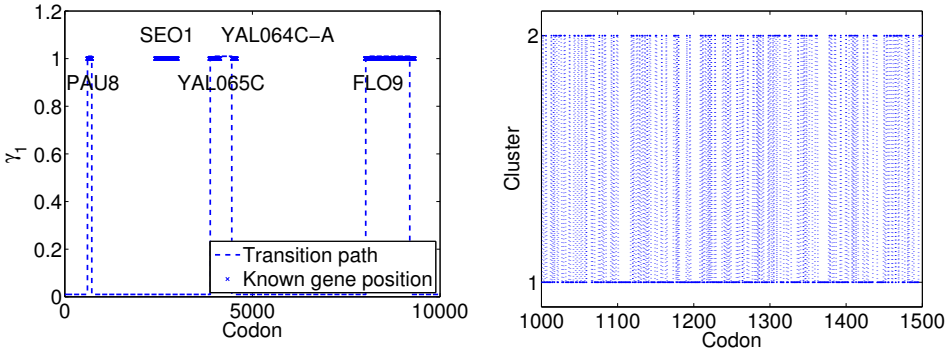


Figure 12. Yeast DNA. Left: a comparison of $\gamma_1(t)$ with the positions of some genes justifies the interpretation of the two optimal clusters as coding and noncoding models. Right: the first part of a Viterbi path of an unsupervised two-state HMM fitting reveals that traditional methods assuming stationarity fail to capture the inherent persistence in codon sequences.

strategies. In contrast, respecting the inherent persistence in the sequence of the codons via the variational FEM-BV approach allowed us to identify most of the known gene positions. Even more important, the detection of coding and noncoding regions, i.e., $\boldsymbol{K}^* = 2$, was part of the result and not a priori included knowledge.

5.e. ***Financial data for commodities.*** In the final example, a time series of daily closing prices of futures on oil is analyzed in order to address two important question: Does the FEM-BV approach allow us to identify market phases (e.g., economic crises) and how do external factors affect the evolution of financial data. In the remainder of the example, we apply the prediction scheme introduced in Section 4 and compare its prediction skills with those of simple prediction methods.

In 1989 J. Hamilton [27] introduced a numerical method to identify what he called hidden market phases in financial data which can be seen as the first combination of nonstationary time series analysis and mathematical finance. Since then the method has been generalized and extended to multidimensional data. Prominent phase-identification techniques are based, e.g., on linear vector autoregressive (VAR) models [48], wavelets [2], Kalman filters [45], (G)ARCH [15; 7] or perfect knowledge about the hidden process [10]. These methods, however, suffer from infeasible numerical complexity in high dimensions (curse of dimensions) or are based on strong model assumptions on the underlying dynamics, e.g., stationarity or Markovianity.

The time series (x_t) under consideration here consists of daily closing prices of futures on the commodity oil for the time horizon 2005–2009 [73]. Futures are very sensitive to changes in market phases because they are broadly traded on speculative reasons. The graph of prices is illustrated in Figure 13. Despite the noisy fluctuations of the daily prices, one can clearly see two tendencies or market phases.

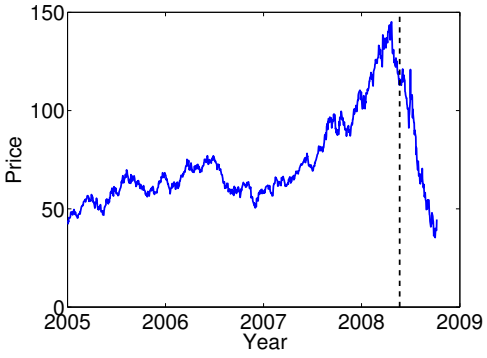


Figure 13. Commodities. Price of oil futures for the timeframe 2005 to 2009. The first 90% of the time series (indicated by the horizontal dashed line) is used as a training set for computing the optimal substitute model. The prediction skill of the nonstationary prediction scheme derived in Section 4 is then assessed on the remaining 10%.

Recall that we are interested in detecting market phases and, more importantly, how their dynamics are affected by external factors. As explained in Section 2.c.iv, the FEM-BV-Markov lends itself well to answer the questions since it allows us to incorporate external factors, specifically.

To this end, the time series of daily prices (x_t) is coarse grained by assigning (x_t) to one of the following categories: (i) The price increased significantly, (ii) no major movement was detected or (iii) the price dropped by a significant amount. Formally, we label the continuous prices by

$$s_t \overset{\text{def}}{=} \begin{cases} 1 & \text{if } x_t - x_{t-1} > \xi, \\ -1 & \text{if } x_t - x_{t-1} < -\xi, \\ 0 & \text{otherwise,} \end{cases} \tag{125}$$

where the threshold ξ separates noise from significant changes and was set to the standard deviation of the time series. This data preparation approach is similar to the one introduced in [27] to detect changes in the Markovian market dynamics.

The transformed time series $(s_t),\ t = 0, \ldots, T$ now takes values in the discrete state space $\mathcal{S} = \{-1, 0, 1\}$. Analogously to the proceeding in Section 2.c.iii, we represent a state s_t by a Dirac-distribution π_t which is defined for the discrete states $s = -1, 0, 1$ as (cf. (30))

$$\pi_t(s) \overset{\text{def}}{=} \begin{cases} 1 & \text{if } s = s_t, \\ 0 & \text{otherwise.} \end{cases} \tag{126}$$

The resulting time series $(\pi_t),\ t = 0, \ldots, T$ encoding the inherent tendencies of the price evolution in terms of probability distributions can now be analyzed by the Markov regression model in (39). More importantly, the FEM-BV-Markov approach allows us to investigate the influence of external factors. Specifically, we would like to understand to which extent the price evolution is influenced by the overall state of the US economy and the climate situation, especially, by the effects of El Niño and La Niña [72].

To this end, the following external factors are considered:

u_1 the daily closing value of the Dow Jones Industrial Average (available at [75])

u_2 the El Niño-Southern Oscillation (ENSO) index 3.4 (available at [47]).

To test on memory effects, three additional external factors are taken into account:

u_3 the Dow Jones shifted (delayed) by one day,

u_4 the ENSO index delayed by 30 days,

u_5 and the ENSO index delayed by 60 days.

Finally, the external factors are scaled to the interval $[0, 1]$ to ensure comparability of the influences as the Dow Jones takes values around $10,000$ while the ENSO takes values between ± 1.5.

Besides the analysis of the data, the main goal of this example is to demonstrate the skills of the prediction scheme presented in Section 4. Therefore, the time series (π_t) is divided into a training set, containing the first 90% of the data and a prediction set, consisting of the remaining data. The analysis via FEM-BV-Markov is based only on the training set, simulating the lack of knowledge about the future, so that the prediction can then be compared to the prediction set.

Next, we describe in detail the clustering of the training set via the FEM-BV-Markov approach and the subsequent optimal model selection. We globally minimized the average cluster functional resulting from the model distance function in (42) for all combinations of $\boldsymbol{K} \in \{1, \ldots, 4\}$ and $\boldsymbol{C} \in \{3, \ldots, 10\}$ and all 2^5 possible subsets of combinations of external factors (ranging from no external factor to all five factors).

Analogously to the proceeding in the previous example in Section 5.d, we exploit the fact that the average mixture model associated with the FEM-BV-Markov approach (cf. Section 2.g and (106)),

$$\hat{\pi}_{t+1}^{\dagger} = \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) \pi_t^{\dagger} P^{(i)}(u(t)), \tag{127}$$

preserves probability, i.e., $\hat{\pi}_{t+1}$ is again a probability distribution. Consequently, the likelihood function $\mathscr{L}(\boldsymbol{K}, \boldsymbol{C})$ (involved in the modified AIC value (97)), here can be computed via (127) by

$$\mathscr{L}(\boldsymbol{K}, \boldsymbol{C}) = \prod_{t=0}^{T-1} \mathbb{P}_{\hat{\pi}_{t+1}}[s_{t+1}] = \prod_{t=0}^{T-1} \hat{\pi}_{t+1}(s_{t+1}). \tag{128}$$

The overall minimum of the modified AIC value with respect to all combinations of clusters' numbers, persistence values and all combination of external factors is attained for $\boldsymbol{K}^* = 2$, $\boldsymbol{C}^* = 6$ and without any external factors. That outcome is consistent with the weak efficient-market hypothesis in [16], stating that any information publicly available is instantly included in the price. The associated affiliation vector (depicted in Figure 14) more or less separates the time horizon of the training data set into two persistent regions. Interestingly, the time point of change at the end of 2008 from cluster 1 to cluster 2 is very close to the beginning of the financial crisis of the late 2000s.

The interpretation of $\Gamma^*(t)$ as an indicator of market phases is further substantiated by looking at the constant transition matrices associated with the two clusters

$$P_0^{(*1)} = \begin{bmatrix} 0.0448 & \mathbf{0.8955} & 0.0597 \\ 0.0989 & \mathbf{0.8112} & 0.0899 \\ 0.1167 & \mathbf{0.8000} & 0.0833 \end{bmatrix}, \quad P_0^{(*2)} = \begin{bmatrix} 0.2453 & 0.4528 & 0.3019 \\ 0.4030 & 0.3433 & 0.2537 \\ 0.3333 & 0.3778 & 0.2889 \end{bmatrix}. \tag{129}$$

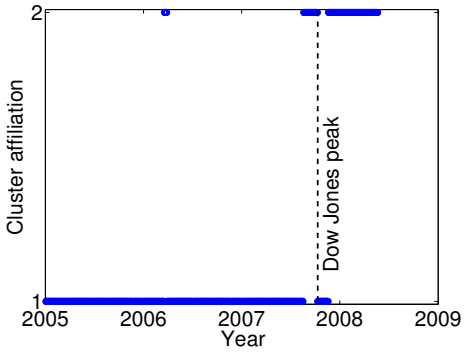


Figure 14. Commodities. The cluster affiliation $\gamma_1^*(t)$ associated with the optimal substitute model with $\boldsymbol{K}^* = 2$, $\boldsymbol{C}^* = 6$ and no external factors for the training set (first 90% of the data). The majority of the second cluster is located from the end of 2007 onwards, indicating a relation to the financial crisis.

Recalling that an entry $P_{ij},\ i, j \in \{-1, 0, 1\}$ of stochastic matrix P with respect to to $\mathcal{S}$ denotes the conditional probability that the associated Markov chain jumps from state i to state j, the second column in $P_0^{(*1)}$ indicates that the noise state $s = 0$ is metastable. In other words, cluster (market phase) $i = 1$ is characterized by small movements without any specific tendencies. In contrast, the transition matrix $P_0^{(*2)}$ of the second cluster does not show any dominating state as the transition probabilities are close to each other, thus, indicating no specific direction in price movement. Additionally, the second column suggests that the average change in price is increased compared to the first cluster. Both observations together imply an increase of the variance in the price evolution which is consistent with the observations in [6; 15] stating that economic crises are characterized by high variance whereas low-variance phases correspond to the normal state of the market.

The analysis was performed for different ending times of the training set, though a relevant influence of the external factors could not be observed. However, if the training set does not include the peak in the price, the analysis yields in selecting the stationary ($\boldsymbol{K}^* = 1$) model. This is to be expected, as the second cluster, representing the "crisis state", has insufficient size to be statistically relevant.

The remainder of this section is devoted to the prediction scheme introduced in Section 4. Rather than predicting the price evolution, we adapt the scheme for predicting the probability distributions $\hat{\pi}_t$ with respect to the discrete state space $\mathcal{S}$ for $t \geq T + 1$.

The fitting scheme associated with the optimal model ($\boldsymbol{K}^* = 2$, $\boldsymbol{C}^* = 6$ and without any external factors) reduces to

$$\hat{\pi}_{t+1}^{\dagger} = \sum_{i=1}^{2} \gamma_i^*(t) \pi_t^{\dagger} P_0^{(i)} \quad t = 0, \ldots, T, \tag{130}$$

where $P_0^{(*1)}$ and $P_0^{(*2)}$ are given in (129). In order to extend (130) to $t \geq T+1$, we estimated a stationary Markov regression model $P^*(u(t))$ based for the time series $(\Gamma^*(t))$ of optimal affiliation vectors. Consistently with the analysis of (π_t), we thereby considered all combinations of external factors. It turned out that the optimal stationary Markov regression model is independent of any external factors too. Formally, we have $P^*(u(t)) = P^*$ and the prediction scheme for $\hat{\Gamma}(t)$ takes the form

$$\hat{\Gamma}^{\dagger}_{[0,T+d]}(T+r) = (\Gamma^*_{[0,T]})^{\dagger}(T)\left[P^*\right]^r, \quad r = 1, \ldots, d. \tag{131}$$

Combining (131) with (127) defines a self-contained nonstationary online prediction scheme analogously to the scheme given in (113). We compare our scheme with standard prediction schemes based on:

(1) An independent stationary model formally given by

$$\hat{\pi}^0_{t+1} = \mu, \quad \mu(s) \stackrel{\text{def}}{=} \frac{1}{T+1}\sum_{t=0}^{T} \chi_s(s_t), \ s \in \{-1, 0, 1\}. \tag{132}$$

(2) A stationary Markov regression model estimated from the time series (π_t) (without any external factors),

$$\left(\hat{\pi}^1_{t+1}\right)^{\dagger} = \left(\hat{\pi}^1_t\right)^{\dagger} P. \tag{133}$$

(3) A zero-prediction model, where the prediction is the last known state

$$\hat{\pi}^{(2)}_{t+1} = \hat{\pi}^{(2)}_t. \tag{134}$$

(4) An artificial neural network, as used in pattern recognition (see [5], for instance), using the external factors as input variables. For the test we have chosen the optimal network configuration (i.e., number of hidden neurons, transfer functions, etc.) with respect to prediction quality.

The average relative prediction error of the five d-step prediction schemes for a prediction horizon $[T+1, T+T']$ on the prediction set is measured similar to (116):

$$\bar{e}_d(T') \stackrel{\text{def}}{=} \frac{1}{T'-d+1}\sum_{t'=T}^{T'-d} \frac{\|x_{t'+d} - \hat{x}_{t'+d}\|_2}{\sqrt{2}}, \tag{135}$$

where the additional factor $\frac{1}{\sqrt{2}}$ is introduced to normalize the error for the worst prediction (the prediction of one state having probability one that is not fitting the realization) to 1. As one can see in Figure 15 (left panel), our nonstationary scheme outperforms the standard schemes. The highly oscillatory behavior of the zero-prediction comes from the fact, that the error of a single prediction is either

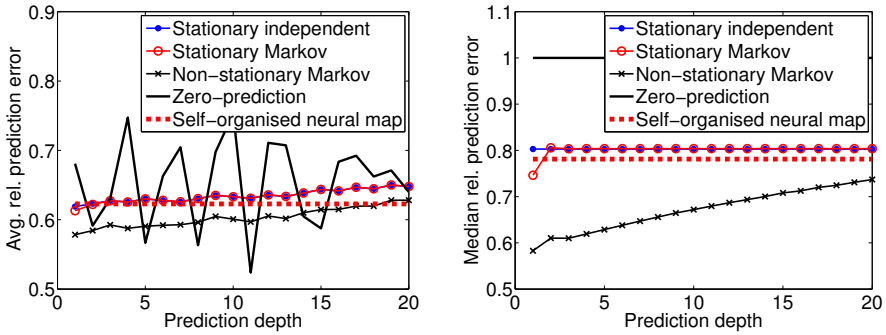


Figure 15. Commodities. The five prediction schemes (see page 220) are compared. Shown are the graphs of the average relative prediction error $\bar{e}_d(T')$ (left) and the median of the relative prediction error (right) as functions of the prediction depth d. They clearly reveal that the nonstationary prediction strategy outperforms the standard schemes.

0 or 1, thus maximizing the variance of the prediction and the small sample size of the predicted time frame. More precisely, when using the median (or 50%-quantile) of the error instead of the average, shown in Figure 15 (right panel), the zero-prediction is more likely wrong than right.

To sum up, we can now answer the questions from the beginning of this section: does the FEM-BV approach allow us to identify market phases (e.g., economic crises) and how do external factors affect the evolution of financial data? First, the FEM-BV-Markov model does not only allow the identification of market phases, but also results in a more accurate model of the market that can be used to predict further movements. Second, in line with the (weak) efficient market hypothesis, the influence of the general US economy, represented by the Dow Jones, and the El Niño/La Niña events were shown the be insignificant with respect to the analyzed data. However, we are aware of the fact, that this might be a result of the insufficient length of the analyzed time series and not a general fact.

We also want to emphasize, that we performed a qualitative analysis instead of a quantitative, as we coarse grained the data to overcome noise effects. While the results might be of no great practical use for investment strategists, it can be considered relevant for risk management, as we were able to verify the fact, that financial unstable market situations yield in a higher volatility. This is a nonnegligible part of most definitions of financial and economical risk.

6. Conclusion

A variational approach to nonstationary time series analysis developed in the last years is presented as a unified framework for analysis, discrimination and prediction of various types of observed processes. It was demonstrated, that persistence

is one of the main characteristic features of many real life processes and that an appropriate mathematical regularization strategy is the clue to its efficient recovery from the observation data. Moreover, a unified model discrimination approach is suggested based on a modified formulation of the information theoretic criterion AIC. Furthermore, the paper contains a first systematic comparison of the FEM-BV methodology with standard time series analysis methods and their underlying mathematical assumptions. The framework is demonstrated on various examples ranging from simple toy models to the analysis of real-world processes such as biomolecular dynamics, DNA-sequence analysis and financial risk prediction.

The effect of nonstationarity is captured in the FEM-BV approach by identifying a (hidden) process in parameter space describing transitions between different regimes which are characterized by local models and their stationary parameters. The presented clustering scheme involves several numerical optimization techniques combining elements of convex optimization, linear programming and Finite Element methods. This allows the employment of fast and numerically robust solvers which ensure an efficient analysis of high dimensional time series. As demonstrated in the present paper, the variational framework is very flexible with respect to different (non-)dynamical scenarios because only the estimators for the optimal parameters have to be provided either analytically or numerically whereas the estimation of the transition process remains general. Therefore, the FEM-BV approach can be straightforwardly adapted and redesigned to new model functions and new applications.

In contrast to classical methods such as HMM, GMM, neuronal networks or local kernel methods, the approach presented here does not rely on a priori probabilistic assumptions (e.g., stationarity, independence, Gaussianity, Markovianity, etc.). Instead of the probabilistic assumptions made in standard statistical methods, here firstly it is assumed that the dynamics under consideration are persistent, i.e., the parameters of the process vary much more slowly than the process itself. Secondly, it is assumed that the hidden process in parameter space can be described by a function with bounded variation. The latter assumption leads to a direct control of the regularity of the hidden process within the course of optimization and, thus, allows us to explicitly incorporate persistence or metastability. For the nonregularized case, it was demonstrated that standard methods such as k-means or (time-dependent) probabilistic mixture models are recovered by the FEM-BV approach as special cases. Although these assumptions are quite general, it is important to emphasize that their fulfillment is crucial for postprocessing and interpretation of the obtained results.

Another aim of this paper was to present a novel self-contained model selection strategy to simultaneously identify the optimal number of clusters and the optimal regularity of the paths in parameter space. As demonstrated in the numerical

examples, the clusterwise approximation of the scalar residuals via maximum entropy distributions in conjunction with the subsequent evaluation of the modified Akaike information criterion successfully allows us to identify the essential nonstationary patterns in various time series. The maximum entropy ansatz follows the philosophy of the FEM-BV approach in that it requires as less as possible explicit a priori knowledge. The central mathematical assumption underlying this strategy is, however, that the scalar residuals are independent. Hence, further research has to be done to generalize this setting in order to cover the case of dependent residuals by, e.g., fitting scalar regression models by means of the maximum entropy principle.

Furthermore, a unified concept for nonstationary time series prediction is presented. While predicting within the trained time span is reduced to evaluation of mixture models, the construction of predictive models beyond that time span requires the understanding of the underlying (learned) transition process in parameter space. To this end, the process of affiliation vectors interpreted as a time series of discrete probability distributions was approximated in terms of a (single) discrete time Markov chain. Predicting an affiliation vector for $t = T + 1$ then allows the approximation of x_{T+1} via a mixture model and so on. However, we are aware that the resulting self-contained prediction strategy crucially relies on the assumption that the memory depth of the affiliation process is at most one. This issue is also the matter of future research.

Appendix

In the appendix we compactly state the constrained quadratic program characterizing the optimal Markov regression model in the FEM-BV-Markov approach. For details see Section 2.c.iv).

Let $\mathbf{vec}(P_l^{(i)}) \in \mathbb{R}^{M^2}$ be the vector which results from concatenating all columns of the matrix $P_l^{(i)}$, i.e.,

$$\mathbf{vec}(P_l^{(i)}) \stackrel{\text{def}}{=} \left(P_l^{(i)}(\cdot, 1), \ldots, P_l^{(i)}(\cdot, M)\right) \in \mathbb{R}^{M^2}. \tag{136}$$

Furthermore, we denote the concatenation of all matrices $P_l^{(i)},\ l = 0, \ldots, k$ as

$$\boldsymbol{p}^{(i)} \stackrel{\text{def}}{=} \left(\mathbf{vec}(P_0^{(i)}), \ldots, \mathbf{vec}(P_k^{(i)})\right) \in \mathbb{R}^{(k+1)M^2}. \tag{137}$$

If we define

$$\boldsymbol{b}^{(i)} = -2 \sum_{t=0}^{T-1} \gamma_i(t) b(t) \quad \text{and } \boldsymbol{H}^{(i)} = 2 \sum_{t=0}^{T-1} \gamma_i(t) H(t) \tag{138}$$

with $u_0(t) \equiv 1$,

$$b(t) = \left(u_0(t)\mathbf{vec}(\pi_t \pi_{t+1}^{\dagger}), \ldots, u_k(t)\mathbf{vec}(\pi_t \pi_{t+1}^{\dagger})\right) \in \mathbb{R}^{(k+1)M^2} \tag{139}$$

and $H(t) \in \mathbb{R}^{((k+1)M^2)\times((k+1)M^2)}$ consists of blocks $H_{l_1 l_2}(t),\ l_1, l_2 = 0, \ldots, k$ with

$$H_{l_1 l_2}(t) = u_{l_1}(t) u_{l_2}(t) \mathrm{diag}(\pi_t \pi_t^\dagger, \ldots, \pi_t \pi_t^\dagger) \in \mathbb{R}^{M^2 \times M^2} \tag{140}$$

then for fixed Γ the solution of the variational problem with respect to i-th local stationary Markov model $\Theta^{(i)} = (P_0^{(i)}, \ldots, P_k^{(i)})$,

$$\boldsymbol{L}(\Theta^{(i)}, \Gamma) = \sum_{t=0}^{T-1} \sum_{i=1}^{\boldsymbol{K}} \gamma_i(t) \left\| {\pi_{t+1}}^\dagger - {\pi_t}^\dagger \left(P_{(0)}^{(i)} + \sum_{l=1}^{k} u_l(t) P_{(l)}^{(i)} \right) \right\|_2^2 \to \min_{\Theta^{(i)}}, \tag{141}$$

subject to the constraints (45)–(48) is given by the solution of

$$\boldsymbol{L}(\boldsymbol{p}^{(i)}) = \tfrac{1}{2} \left\langle \boldsymbol{p}^{(i)}, \boldsymbol{H}^{(i)} \boldsymbol{p}^{(i)} \right\rangle_2 + \left\langle \boldsymbol{b}^{(i)}, \boldsymbol{p}^{(i)} \right\rangle_2 \to \min_{\boldsymbol{p}^{(i)}} \tag{142}$$

subject to the following linear constraints:

- Nonnegativity constraints (45):

$$\underbrace{(\mathrm{Id}_{M^2}, 0, \ldots, 0)}_{\in \mathbb{R}^{M^2 \times (k+1)M^2}} \boldsymbol{p}^{(i)} \geq 0, \tag{143}$$

- Constraints (46) and (47):

$$\underbrace{\begin{pmatrix} \mathcal{R}(\mathbf{1}_M) & 0 & 0 & 0 \\ 0 & \mathcal{R}(\mathbf{1}_M) & 0 & 0 \\ & & \ddots & \\ 0 & 0 & 0 & \mathcal{R}(\mathbf{1}_M) \end{pmatrix}}_{\in \mathbb{R}^{(k+1)M \times (k+1)M^2}} \boldsymbol{p}^{(i)} = \begin{pmatrix} \mathbf{1}_M \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{144}$$

  with $\mathcal{R}(\mathbf{1}_M) = (\mathrm{Id}_M, \ldots, \mathrm{Id}_M) \in \mathbb{R}^{M \times M^2}$.

- Overall nonnegativity constraint in (48):

$$\underbrace{\left(\mathrm{Id}_{M^2}, \hat{u}_1 \mathrm{Id}_{M^2}, \ldots, \hat{u}_k \mathrm{Id}_{M^2}\right)}_{\in \mathbb{R}^{M^2 \times (k+1)M^2}} \boldsymbol{p}^{(i)} \geq 0 \tag{145}$$

  for all $(\hat{u}_1, \ldots, \hat{u}_k) \in \{a_1, b_1\} \times \ldots \times \{a_k, b_k\}$ with

$$a_l = \min\{u_l(t) : t = 0, \ldots, T\} \quad \text{and} \quad b_l = \max\{u_l(t) : t = 0, \ldots, T\}. \tag{146}$$

Acknowledgement

We would like to thank the anonymous referees for their constructive criticism which helped us to improve the readability of the manuscript.

References

[1] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Automat. Control **19** (1974), no. 6, 716–723. MR 54 #11691 Zbl 0314.62039

[2] A. N. Akansu and R. A. Haddad, *Multiresolution signal decomposition: transforms, subbands, and wavelets*, Academic Press, Boston, 1992. MR 93m:94004 Zbl 0947.94001

[3] L. E. Baum, *An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes*, Inequalities, III (O. Shisha, ed.), Academic Press, New York, 1972, pp. 1–8. MR 49 #6528

[4] L. E. Baum and T. Petrie, *Statistical inference for probabilistic functions of finite state Markov chains*, Ann. Math. Stat. **37** (1966), no. 6, 1554–1563. MR 34 #2137 Zbl 0144.40902

[5] C. M. Bishop, *Neural networks for pattern recognition*, Clarendon, Oxford, 1995. MR 97m: 68172 Zbl 0868.68096

[6] F. Black, *Studies of stock price volatility changes*, Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economics Statistics Section, American Statistical Association, Washington, DC, 1976, pp. 177–181.

[7] T. Bollerslev, *Generalized autoregressive conditional heteroskedasticity*, J. Econometrics **31** (1986), no. 3, 307–327. MR 87j:62169 Zbl 0616.62119

[8] D. Braess, *Finite elements: theory, fast solvers, and applications in solid mechanics*, 2nd ed., Cambridge University Press, Cambridge, 2001. MR 2001k:65002 Zbl 0976.65099

[9] P. Brémaud, *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, Texts in Applied Mathematics, no. 31, Springer, New York, 1999. MR 2000k:60137 Zbl 0949.60009

[10] U. Çelikyurt and S. Özekici, *Multiperiod portfolio optimization models in stochastic markets using the mean-variance approach*, Eur. J. Oper. Res. **179** (2007), no. 1, 186–202. Zbl 1163.91375

[11] M. Dellnitz and O. Junge, *On the approximation of complicated dynamical behavior*, SIAM J. Numer. Anal. **36** (1999), no. 2, 491–515. MR 2000c:37026 Zbl 0916.58021

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Stat. Soc. Ser. B **39** (1977), no. 1, 1–38. MR 58 #18858 Zbl 0364.62022

[13] P. Deuflhard and C. Schütte, *Molecular conformation dynamics and computational drug design*, Applied mathematics entering the 21st century (J. M. Hill and R. Moore, eds.), SIAM, Philadelphia, 2004, pp. 91–119. MR 2296264 Zbl 1134.92004

[14] P. Deuflhard and M. Weber, *Robust Perron cluster analysis in conformation dynamics*, Linear Algebra Appl. **398** (2005), 161–184. MR 2005h:62166 Zbl 1070.15019

[15] R. F. Engle, *Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation*, Econometrica **50** (1982), no. 4, 987–1007. MR 83j:62158 Zbl 0491. 62099

[16] E. F. Fama, *The behavior of stock-market prices*, J. Bus. **38** (1965), no. 1, 34–105.

[17] A. Fischer, S. Waldhausen, I. Horenko, E. Meerbach, and C. Schütte, *Identification of biomolecular conformations from incomplete torsion angle observations by hidden Markov models*, J. Comput. Chem. **28** (2007), no. 15, 2453–2464.

[18] C. Franzke, D. Crommelin, A. Fischer, and A. J. Majda, *A hidden Markov model perspective on regimes and metastability in atmospheric flows*, J. Climate **21** (2008), no. 8, 1740–1757.

[19] T. Gasser and H.-G. Müller, *Kernel estimation of regression functions*, Smoothing techniques for curve estimation (T. Gasser and M. Rosenblatt, eds.), Lecture Notes in Math., no. 757, Springer, Berlin, 1979, pp. 23–68. MR 81k:62052 Zbl 0418.62033

[20] ———, *Estimating regression functions and their derivatives by the kernel method*, Scand. J. Stat. **11** (1984), no. 3, 171–185. MR 86h:62056 Zbl 0548.62028

[21] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, 2nd ed., Chapman & Hall/CRC, Boca Raton, FL, 2004. MR 2004j:62001 Zbl 1039.62018

[22] D. Giannakis and A. J. Majda, *Quantifying the predictive skill in long-range forecasting, I: Coarse-grained predictions in a simple ocean model*, J. Climate **25** (2011), 1793–1813.

[23] ———, *Quantifying the predictive skill in long-range forecasting, II: Model error in coarse-grained Markov models with application to ocean-circulation regimes*, J. Climate **25** (2011), 1814–1826.

[24] D. Giannakis, A. J. Majda, and I. Horenko, *Information theory, model error, and predictive skill of stochastic models for complex nonlinear systems*, preprint, CIMS/NYU and University of Lugano, 2011, Submitted to Physica D.

[25] A. L. Gibbs and F. E. Su, *On choosing and bounding probability metrics*, Int. Stat. Rev. **70** (2002), no. 3, 419–435. Zbl 1217.62014

[26] J. Hadamard, *Sur les problèmes aux dérivées partielles et leur signification physique*, Princeton Univ. Bull. **13** (1902), 49–52.

[27] J. D. Hamilton, *A new approach to the economic analysis of nonstationary time series and the business cycle*, Econometrica **57** (1989), no. 2, 357–384. MR 996941 Zbl 0685.62092

[28] A. Hoerl, *Application of ridge analysis to regression problems*, Chem. Eng. Prog. **58** (1962), no. 3, 54–59.

[29] I. Horenko, *On simultaneous data-based dimension reduction and hidden phase identification*, J. Atmos. Sci. **65** (2008), no. 6, 1941–1954.

[30] ———, *On robust estimation of low-frequency variability trends in discrete Markovian sequences of Atmospherical Circulation Patterns*, J. Atmos. Sci. **66** (2009), no. 7, 2059–2072.

[31] ———, *Finite element approach to clustering of multidimensional time series*, SIAM J. Sci. Comput. **32** (2010), no. 1, 62–83. MR 2011b:62009 Zbl 1206.62150

[32] ———, *On clustering of non-stationary meteorological time series*, Dyn. of Atmos. and Oceans **49** (2010), no. 2-3, 164–187.

[33] ———, *On identification of non-stationary factor models and its application to atmospherical data analysis*, J. Atmos. Sci. **67** (2010), no. 5, 1559–1574.

[34] ———, *Nonstationarity in multifactor models of discrete jump processes, memory and application to cloud modeling*, J. Atmos. Sci. **68** (2011), no. 7, 1493–1506.

[35] ———, *On analysis of nonstationary categorical data time series: dynamical dimension reduction, model selection, and applications to computational sociology*, Multiscale Model. Simul. **9** (2011), no. 4, 1700–1726. MR 2861255

[36] I. Horenko, E. Dittmer, A. Fischer, and C. Schütte, *Automated model reduction for complex systems exhibiting metastability*, Multiscale Model. Simul. **5** (2006), no. 3, 802–827. MR 2257236 Zbl 1122.60062

[37] I. Horenko, E. Dittmer, and C. Schütte, *Reduced stochastic models for complex molecular systems*, Comput. Vis. Sci. **9** (2006), no. 2, 89–102. MR 2247687

[38] I. Horenko, S. Dolaptchiev, A. Eliseev, I. Mokhov, and R. Klein, *Metastable decomposition of high-dimensional meteorological data with gaps*, J. Atmos. Sci. **65** (2008), no. 11, 3479–3496.

[39] I. Horenko, R. Klein, S. Dolaptchiev, and C. Schütte, *Automated generation of reduced stochastic weather models, I: Simultaneous dimension and model reduction for time series analysis*, Multiscale Model. Simul. **6** (2007), no. 4, 1125–1145. MR 2009e:62338 Zbl 1152.62056

[40] I. Horenko, J. Schmidt-Ehrenberg, and C. Schütte, *Set-oriented dimension reduction: localizing principal component analysis via hidden Markov models*, Computational life sciences II (M. R. Berthold, R. Glen, and I. Fischer, eds.), Lecture Notes in Comput. Sci., no. 4216, Springer, Berlin, 2006, pp. 74–85. MR 2279311

[41] I. Horenko and C. Schütte, *Likelihood-based estimation of multidimensional Langevin models and its application to biomolecular dynamics*, Multiscale Model. Simul. **7** (2008), no. 2, 731–773. MR 2009j:37136 Zbl 1180.35175

[42] W. Huisinga, *Metastability of Markovian systems: a transfer operator approach in application to molecular dynamics*, Ph.D. thesis, Free University Berlin, 2001.

[43] E. T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev. (2) **106** (1957), 620–630. MR 19,335b Zbl 0084.43701

[44] ———, *Information theory and statistical mechanics, II*, Phys. Rev. (2) **108** (1957), 171–190. MR 20 #2898 Zbl 0084.43701

[45] R. Kalman, *A new approach to linear filtering and prediction problems*, J. Basic Eng. **82** (1960), no. 1, 35–45.

[46] J. N. Kapur, *Maximum-entropy models in science and engineering*, Wiley, New York, 1989. MR 92b:00017 Zbl 0746.00014

[47] Koninklijk Nederlands Meteorologisch Instituut, http://climexp.knmi.nl/getindices.cgi?WMO=NCEPData/nino2_daily&STATION=NINO12&id=someone@somewhere&NPERYEAR=366&TYPE=i.

[48] H.-M. Krolzig, *Predicting Markov-switching vector autoregressive processes*, preprint 2000-W31, University of Oxford, 2000.

[49] D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman, *A generalized hidden Markov model for the recognition of human genes in DNA*, Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology (D. J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. Smith, eds.), AAAI Press, Menlo Park, CA, 1996, pp. 134–142.

[50] C. L. Lawson and R. J. Hanson, *Solving least squares problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974. MR 51 #2270 Zbl 0860.65028

[51] C. Lee and M.-H. Yu, *Protein folding and disease*, J. Biochem. Molec. Biol. **38** (2005), no. 3, 275–280.

[52] C. Loader, *Local regression and likelihood*, Springer, New York, 1999. MR 2000f:62005 Zbl 0929.62046

[53] J. B. MacQueen, *Some methods for classification and analysis of multivariate observations*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, I: Statistics (L. M. Le Cam and J. Neyman, eds.), University of California Press, Berkeley, CA, 1967, pp. 281–297. MR 35 #5078 Zbl 0214.46201

[54] A. J. Majda, C. L. Franzke, A. Fischer, and D. T. Crommelin, *Distinct metastable atmospheric regimes despite nearly Gaussian statistics: a paradigm model*, Proc. Natl. Acad. Sci. USA **103** (2006), no. 22, 8309–8314. MR 2007a:86004 Zbl 1160.86304

[55] A. J. Majda and X. Wang, *Non-linear dynamics and statistical theories for basic geophysical flows*, Cambridge University Press, Cambridge, 2006. MR 2009e:76214 Zbl 1141.86001

[56] G. McLachlan and D. Peel, *Finite mixture models*, Wiley, New York, 2000. MR 2002b:62025 Zbl 0963.62061

[57] P. Metzner, M. Weber, and C. Schütte, *Observation uncertainty in reversible Markov chains*, Phys. Rev. E (3) **82** (2010), no. 3, Paper #031114. MR 2012a:60202

[58] J.-J. Moreau, P. D. Panagiotopoulos, and G. Strang (eds.), *Topics in nonsmooth mechanics*, Birkhäuser, Basel, 1988.

[59] National Center for Biotechnology Information, *Saccharomyces cerevisiae chromosome I, complete sequence*, http://www.ncbi.nlm.nih.gov/nuccore/144228165?report=graph&to=30000.

[60] R. Preis, M. Dellnitz, M. Hessel, C. Schütte, and E. Meerbach, *Dominant paths between almost invariant sets of dynamical systems*, preprint 154, Deusche Forschungsgemeinschaft Schwerpunktprogramm, 2004.

[61] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes: the art of scientific computing*, 3rd ed., Cambridge University Press, Cambridge, 2007. MR 2009b:65001 Zbl 1132.65001

[62] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, *Markov models of molecular kinetics: generation and validation*, J. Chem. Phys. **134** (2011), no. 17, Paper #174105.

[63] L. Putzig, D. Becherer, and I. Horenko, *Optimal allocation of a futures portfolio utilizing numerical market phase detection*, SIAM J. Financial Math. **1** (2010), 752–779. MR 2011k:91171 Zbl 1198.91241

[64] L. R. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition*, Proc. IEEE **77** (1989), no. 2, 257–286.

[65] C. Schütte, *Conformational dynamics: modelling, theory, algorithm, and application to biomolecules*, preprint SC 99-18, Konrad-Zuse-Zentrum für Informationstechnik, Berlin, 1999.

[66] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, *A direct approach to conformational dynamics based on hybrid Monte Carlo*, J. Comput. Phys. **151** (1999), no. 1, 146–168. MR 2000d: 92004

[67] C. Schütte and W. Huisinga, *Biomolecular conformations can be identified as metastable sets of molecular dynamics*, Handbook of numerical analysis, 10: Computational chemistry (C. Le Bris, ed.), North-Holland, Amsterdam, 2003, pp. 699–744. MR 2008396 Zbl 1066.81658

[68] C. Schütte, F. Noé, J. Lu, M. Sarich, and E. Vanden-Eijnden, *Markov state models based on milestoning*, J. Chem. Phys. **134** (2011), no. 20, Paper #204105.

[69] C. Schütte, F. Noé, E. Meerbach, P. Metzner, and C. Hartmann, *Conformation dynamics*, ICIAM 07: 6th International Congress on Industrial and Applied Mathematics (R. Jeltsch and G. Wanner, eds.), European Mathematical Society, Zürich, 2009, pp. 297–335. MR 2011k:82072 Zbl 1180.82239

[70] F. Takens, *Detecting strange attractors in turbulence*, Dynamical systems and turbulence, Warwick 1980 (D. A. Rand and L.-S. Young, eds.), Lecture Notes in Math., no. 898, Springer, Berlin, 1981, pp. 366–381. MR 83i:58065 Zbl 0513.58032

[71] A. N. Tikhonov, *On the stability of inverse problems*, Dokl. Akad. Nauk SSSR **39** (1943), no. 5, 195–198, In Russian; translated in C. R. (Doklady) Acad. Sci. URSS (N.S.) **39** (1943), no. 5, 176–179. MR 5,184e Zbl 0061.23308

[72] K. E. Trenberth, *The definition of El Niño*, Bull. Amer. Meteorol. Soc. **78** (1997), no. 12, 2771–2777.

[73] U.S. Energy Information Administration, *NYMEX futures prices*, http://tonto.eia.doe.gov/dnav/pet/pet_pri_fut_s1_d.htm.

[74] G. Wahba, *Spline models for observational data*, CBMS-NSF Regional Conference Series in Applied Mathematics, no. 59, SIAM, Philadelphia, 1990. MR 91g:62028 Zbl 0813.62001

[75] Yahoo Finance, *Dow Jones industrial average: historical prices*, http://finance.yahoo.com/q/hp?s=^DJI+Historical+Prices.

[76] A. Zellner and R. A. Highfield, *Calculation of maximum entropy distributions and approximation of marginal posterior distributions*, J. Econometrics **37** (1988), no. 2, 195–209. MR 932140

Received July 29, 2011. Revised March 23, 2012.

PHILIPP METZNER: metznerp@usi.ch
Institute of Computational Science, Faculty of Informatics, Università della Svizzera italiana, Via Guiseppe Buffi 13, CH-6900 Lugano, Switzerland
http://icsweb.inf.unisi.ch/cms/index.php/people/24-philipp-metzner.html

LARS PUTZIG: putzigl@usi.ch
Institute of Computational Science, Faculty of Informatics, Università della Svizzera italiana, Via Guiseppe Buffi 13, CH-6900 Lugano, Switzerland
http://icsweb.inf.unisi.ch/cms/index.php/people/27-lars-putzig.html

ILLIA HORENKO: horenkoi@usi.ch
Institute of Computational Science, Faculty of Informatics, Università della Svizzera italiana, Via Guiseppe Buffi 13, CH-6900 Lugano, Switzerland
http://icsweb.inf.unisi.ch/cms/index.php/people/20-illia-horenko.html

COMM. APP. MATH. AND COMP. SCI.
Vol. 7, No. 2, 2012
dx.doi.org/10.2140/camcos.2012.7.231

msp

APPROXIMATION OF PROBABILISTIC LAPLACE TRANSFORMS AND THEIR INVERSES

GUILLAUME COQUERET

We present a method to approximate the law of positive random variables defined by their Laplace transforms. It is based on the study of the error in the Laplace domain and allows for many behaviors of the law, both at 0 and infinity. In most cases, both the Kantorovich/Wasserstein error and the Kolmogorov–Smirnov error can be accurately computed. Two detailed examples illustrate our results.

1. Introduction

The topic of Laplace transform inversion is an old problem which relates to physics, probability theory, analysis and numerical methods. The number of publications dedicated to it is so large that it is possible to write surveys of surveys on the subject; see [5, Chapter 9]. If f is a positive integrable function on $\mathbb{R}_+$, we define the Laplace transform operator as follows,

$$\mathbb{L}[f(x)](t) = L(t) = \int_0^\infty e^{-tx} f(x)\, ds,$$

When L is given, the inverse Laplace transform operator $\mathbb{L}^{-1}$ applied to L yields the original function f. Two of the most important results related to $\mathbb{L}^{-1}$ are the *Bromwich integral* (see section 2.2 in [5]):

$$\mathbb{L}^{-1}[L(t)](x) = f(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{xt} L(t)\, dt, \tag{1-1}$$

for some c chosen so that the path of integration makes sense for L, and the *Post–Widder formula*: (see [8], section VII.6):

$$f(x) = \lim_{n\to+\infty} \frac{(-1)^n}{n!} \left(\frac{n}{x}\right)^{n+1} L^{(n)}(n/x). \tag{1-2}$$

However, there are many other ingenious ways to obtain f from L. Most techniques of Laplace transform inversion belong to one of the five families of methods listed in Table 1.

MSC2010: primary 65R32; secondary 65C50.
Keywords: approximation, Laplace transform inversion, completely monotone functions, Kantorovich distance.

	Method based on	References
i)	the Bromwich integral	[4], [26], [5, Chapters 4 and 6]
ii)	the Post–Widder formula	[2], [25], [24], [5, Chapter 7]
iii)	Fredholm equations of the first kind	[7], [22, §12.5-3], [14], [5, Chapter 8]
iv)	rational approximation	[12], [15], [16], [5, Chapter 5]
v)	series expansion	[5, Chapter 3]

Table 1

The references related to these methods are of course far from exhaustive.

All of these approaches lead to numerical approximations. A recent survey on the efficiency of some of these procedures was recently carried out by Masol and Teugels in [18]. The families of techniques can further be categorized into two broader sets:

- **i) + ii) + iii)**: methods for which the initial function, L, is exact, but the inversion approximate (discretization of the integral in (1-1) or choice of a large, but finite, n in (1-2))
- **iii) + iv) + v)**: methods for which the inversion is exact, but the target function is an approximation of the initial Laplace transform

(Some methods from the third family can belong to both sets.)

The approximations stemming from the second set of techniques take the form

$$L(t) \approx \sum_{k=1}^{N} c_k L_k(t) \quad \Longleftrightarrow \quad f(x) \approx \sum_{k=1}^{N} c_k f_k(x) \tag{1-3}$$

The core idea of this paper is to take advantage of the properties of Laplace transforms in probability to choose the L_k (and thus f_k) wisely, depending on some properties of f. We present an iterative procedure which progressively reduces the Kantorovich error induced by the approximation. The main contribution of our approach is that when f is bounded from above, this method provides a uniform maximum for the error made on the cumulative distribution function. Such results are quite rare in the literature, but one reference in a slightly different setting is [23].

This method can be used, for instance, to approximate the law of positive infinitely divisible distributions, which are usually characterized by their Laplace transform.

The paper is organized as follows: in Section 2, we detail some of the properties of f which can be inferred from L and which will be used further on. In Section 3, we detail our method and some error related results, and, lastly, we provide numerical examples in Section 4.

2. Some properties of the density

The aim of this section is to recall a few classical results which show that many properties of f can be derived from a thorough study of L.

We begin with some notations. Throughout the paper, we will consider two positive random variables X and Y with densities f and g, cumulative distribution functions (CDFs) F and G and Laplace transforms L and M respectively. We also denote by $\bar{F}(x) = 1 - F(x)$ and $\bar{G}(x) = 1 - G(x)$ their survival functions. The function L (resp. F) will be the original Laplace transform (resp. CDF) and M (resp. G) its approximation.

For a function $f = f^{(0)}$, $f^{(n)}$ will denote its n-th derivative and in some asymptotic settings, we will write $f(x) \sim g(x)$ for $f(x)/g(x) \to 1$.

Support. The first basic piece of information required to characterize a distribution is its support.

Theorem 2.1. *Let A denote the left point of the support of the positive random variable X. Then if B is the set of real numbers b such that $e^{bt}L(t) = O(1)$ as $t \to \infty$, we have*

$$A = \sup_{b \in B} b$$

Proof. If $A = 0$, then, for any $x < 0$, $e^{xt}L(t) \to 0$ $(t \to +\infty)$. For $x > 0$,

$$e^{xt}L(t) \geq \int_0^x e^{st} f(x-s)\, ds \geq \eta e^{\delta t} \mathrm{Leb}\{s \in [\delta, x] : f(x-s) \geq \eta\} \to \infty, \quad t \to +\infty,$$

where Leb is the Lebesgue measure and $\delta, \eta > 0$ were chosen such that

$$\mathrm{Leb}\{s \in [\delta, x] : f(x-s) \geq \eta\} > 0,$$

which is possible, since $A = 0$. The case $A > 0$ follows by direct translation. □

Another way to obtain the lower bound of the support of X is in fact to compute the limit of $-L'(t)/L(t)$ when $t \to \infty$. Indeed, by Hölder's inequality, $\mathrm{Log}\, L$ is convex, hence L'/L is increasing. Since it is bounded above by zero, it converges to some negative limit. A simple analysis shows that this limit at infinity is in fact $-A$.

In order to find the upper bound of the support of X, we propose a test, based on the following proposition. Note that it is easy to compute $\mathbb{E}[X]$ with the sole knowledge of L, since $\mathbb{E}[X] = -L'(0)$.

Proposition 2.2. *If the positive random variable X is almost surely bounded above by C, then for any $A > 0$ and $\gamma \geq 1$,*

$$L(t) \leq 1 - \mathbb{E}[X^\gamma] \frac{1 - e^{-A}}{A^\gamma} t^\gamma \quad \textit{for all } t \in [0, A/C].$$

Proof. The proof relies on the inequality

$$y^\gamma - x^\gamma \geq e^{-x} y^\gamma - e^{-y} x^\gamma, \quad \gamma \geq 1, \quad 0 < x < y.$$

Setting $x = tX$, $y = A$ and applying the expectation operator yields the result. □

Hence, if $L(t) > 1 - t\,\mathbb{E}[X](1-e^{-A})/A$ in the vicinity of 0, then X is unbounded. The test usually performs better for $A \ll 1$.

In the same spirit, note that Theorem 7(b) in [10] makes it possible to build another test based on $\mathbb{E}[X^\gamma]$ for $\gamma < 1$. Since they depend on the interval $[0, A/C]$, these results make it even possible to derive bounds for C.

By Theorem 2.1 and Proposition 2.2, we will henceforth, without much loss of generality, restrict ourselves to distributions with *supports on the whole positive real line*.

Tail behaviors. This subsection recalls classical Tauberian theorems in probability (see for instance [8, XIII.5]). These results show the strong link that exists between the behavior of f near zero and that of L near infinity and vice-versa. The general form of the de Bruijn exponential Tauberian theorem can be found in [3], Theorem 4.12.9, but we recall below a more peculiar form, derived from Corollary 4.12.6 of the same monograph.

Theorem 2.3. *Let $0 < \gamma < 1$, $\delta \in \mathbb{R}$, $C > 0$ and X a positive random variable. Then,*

$$\log \mathbb{E}[e^{-tX}] \sim -Ct^\gamma (\log(t))^\delta, \quad t \to \infty$$

if and only if

$$\log P[X \leq x] \sim -[C\gamma^\gamma (1-\gamma)^{1-\gamma-\delta} x^{-\gamma} (-\log x)^\delta]^{1/(1-\gamma)}, \quad x \downarrow 0.$$

In a series of papers, Nakagawa provides conditions on L to determine whether a distribution has a heavy or a light tail. We state one of his results below (see [19] and the references therein). For a complex number $z = a + ib$, if $L(z)$ converges for $a > a_0$ and diverges for $a < a_0$, then a_0 is said to be the abscissa of convergence of $L(z)$.

Theorem 2.4. *If a_0 is the abscissa of L such that $-\infty < a_0 < 0$ and a_0 is a pole of L, then*

$$\lim_{x \to \infty} \frac{1}{x} \log P[X > x] = a_0.$$

When the asymptotic behaviors are not exponential but of power form, we can also resort to [3, Corollary 8.1.7] and [8, XIII.5, Theorem 2], which we recall below (with $l(x) = C \log(x)^\beta$).

Proposition 2.5. *For $0 \le \alpha < 1$, $\beta \ge 0$ and $C > 0$, the following are equivalent*:

$$1 - L(t) \sim -Ct^{\alpha} \log(t)^{\beta}, \quad t \downarrow 0.$$
$$1 - F(x) \sim C \frac{\log(x)^{\beta}}{x^{\alpha}\Gamma(1-\alpha)}, \quad x \to \infty.$$

Proposition 2.6. *For $\alpha, \beta \ge 0$ and $C > 0$, the following are equivalent*:

$$L(t) \sim C \frac{\log(t)^{\beta}}{t^{\alpha}}, \qquad t \to \infty.$$
$$F(x) \sim -C \frac{\log(x)x^{\alpha}}{\Gamma(1+\alpha)}, \quad x \downarrow 0.$$

The first result allows one to accurately determine the tail of a distribution when it is very heavy. For other power tail behaviors (when $\alpha > 1$), we refer to Theorem 8.1.6 in [3].

Lastly, we recall the initial value theorem:

$$f(0+) = \lim_{t \to \infty} tL(t).$$

Boundedness. It can be very convenient to know whether a distribution has a bounded density. In order to do so, it is possible to build a test based on the following corollary of the Post–Widder formula.

Lemma 2.7. *A function L, defined on $\mathbb{R}_+$ is the Laplace transform of a probability density bounded above by c if and only if $L(0) = 1$ and*

$$0 \le (-1)^n L^{(n)}(t) \le \frac{cn!}{t^{n+1}}$$

for all $n = 0, 1, \ldots$ and $t > 0$.

3. The approximation method

3.1. ***Introductory remarks.*** We shall henceforth consider a given positive function L defined on $\mathbb{R}_+$, satisfying $L(0) = 1$,

$$(-1)^n L^{(n)}(t) \ge 0 \quad \text{for all } t > 0 \text{ and all } n \ge 0 \tag{3-1}$$

and

$$\lim_{t \to \infty} tL(t) < \infty. \tag{3-2}$$

Any function L for which (3-1) holds is called a complete monotone function. Such functions have the following well-known property (see Theorem 7.11 in [27] for instance).

Theorem 3.1. *A function h is completely monotone on $\mathbb{R}_+$ if and only if it is the Laplace transform of a nonnegative finite Borel measure ν, i.e., if*

$$h(x) = \int_0^\infty e^{-xt} \nu(dt).$$

Therefore, (3-1) and $L(0) = 1$ are necessary and sufficient conditions for probabilistic Laplace transforms.

Our approach is essentially error driven: many inversion techniques do not allow to compute the error made on $f(x)$ or $F(x)$. Some techniques come with error bounds, but these bounds usually increase with x (see for instance (4.61) and (6.19) in [5]). Our method will focus on $\int_0^\infty |F(x) - G(x)|\, dx$, where F is the original CDF and G the approximate one. As we will see, this is an appropriate choice, because when f is bounded, it yields a uniform bound on $|F - G|$, which is a strong result.

The main problem with the L^1 error on F is that it cannot be retrieved from $\left|\int_0^\infty (F(x) - G(x))\, dx\right|$, unless the sign of $F - G$ does not change. Notice that focusing on cumulative distribution functions is critical since it can occur that G is dominated by F on $\mathbb{R}_+$ while this is impossible for two probability densities. The aim of our method is thus to find G as close to F as possible, satisfying $G(x) \geq F(x)$ (or $G(x) \leq F(x)$) for all $x \geq 0$.

This property is connected to a notion called stochastic ordering. We will say that the positive random variable X is less than Y *in the usual stochastic order*, abbreviated i.u.s.o., if

$$1 - F(x) = P[X \geq x] \leq P[Y \geq x] = 1 - G(x) \quad \text{for all } x \geq 0.$$

If X is less than Y i.u.s.o., then an integration by parts yields $L(t) \geq M(t)$ for all $t \geq 0$. Sadly, the converse is not always true. A counter-example is given by the densities $f(x) = \frac{1}{2}(\mathbf{1}_{(0,1)}(x) + \mathbf{1}_{(2,3)}(x))$ and $g(x) = \mathbf{1}_{(1,2)}(x)$. In this case, the CDFs are not ordered, while

$$L(t) = \frac{1 - e^{-t} + e^{-2t} - e^{-3t}}{2t} \geq \frac{e^{-t} - e^{-2t}}{t} = M(t), \quad t \geq 0.$$

In order to make sure that $G(x) \geq F(x)$ for all $x \geq 0$, we will quite logically resort to completely monotone functions. Given L, our aim is to find (or build) another probabilistic Laplace transform M, as close as possible (in some sense) to L and such that

$$t \mapsto \mathbb{L}[G(x) - F(x)](t) = \frac{M(t) - L(t)}{t}$$

is a completely monotone function. Under these conditions, the error made on the cumulative distribution functions will have a constant sign. In fact, this idea can be applied any finite number of times in order to get G as close to F as desired.

Practical implementation. We proceed in two steps.

Step 1. The first step is to find M, a *rough* proxy of L. Inspired by the results on tail behaviors (pages 234 and 235), we propose families of approximants depending on tail behaviors.

If X is light-tailed, then a relevant tool to work with is the gamma distribution. Indeed, its tail is light and it allows for any power behavior near the origin, including $f(0+) > 0$. M and G then have the form

$$M(t) = \frac{a^b}{(a+t)^b}, \quad G(x) = \gamma(b, ax)/\Gamma(b), \quad g(x) = \frac{a^b x^{b-1} e^{-ax}}{\Gamma(b)}, \quad a, b > 0,$$

where $\gamma(\cdot, \cdot)$ is the lower gamma function.

If X has heavy tails, then the choice of the Pareto distribution seems quite straightforward when $f(0+) > 0$. That is,

$$M(t) = ba^b e^{at} t^b \Gamma(-b, at), \quad G(x) = 1 - \frac{a^b}{(a+x)^b}, \quad g(x) = \frac{ba^b}{(a+x)^{b+1}}, \quad a, b > 0,$$

where $\Gamma(\cdot, \cdot)$ is the upper gamma function. In this case, $M(t) \sim ba^{-1}/t$, $t \to \infty$.

If $f(0+) = 0$ (and X is heavy-tailed), we propose the following two choices:

- If f goes slowly to 0 ($x \downarrow 0$), set

$$M(t) = \frac{b(b-1)}{at}(1 - e^{at}(at)^b(at+b)\Gamma(-b, at)), \quad G(x) = 1 - a^{b-1}\frac{a+bx}{(a+x)^b},$$
$$g(x) = \frac{b(b-1)a^{b-1}x}{(a+x)^{b+1}}, \quad a > 0,\ b > 1.$$

- If f goes rapidly to 0, set

$$M(t) = \frac{2}{\Gamma(b)}(at)^{b/2} K_b(2\sqrt{at}), \quad G(x) = \frac{\Gamma(b, a/x)}{\Gamma(b)}, \quad g(x) = \frac{a^b e^{-a/x}}{\Gamma(b) x^{b+1}},$$

  with $a, b > 0$, where $K_\nu(x)$ is the modified Bessel function of the second kind with index ν (see 3.471-9 in [9] for the computation of the Laplace transform). This is a generalization of both the Lévy and the inverse chi-square laws, often referred to as the inverse gamma distribution.

The purpose of the rough proxy is to mimic as well as possible the behavior of L at 0 and/or infinity while satisfying the condition that $\mp(M - L)$ is a completely monotonic function.

Step 2. Without loss of generality, we consider $M - L > 0$. The error made with the rough proxy $N(t) = M(t) - L(t)$ is usually not satisfactory and requires improvement. The trick is thus to find an easily Laplace-inverted minorant μ of

N (i.e., $\mu(x) < N(x)$ for all $x > 0$) such that $(N(t) - \mu(t))/t$ is a completely monotone function. Consequently,

$$G(x) \geq G(x) - \mathbb{L}^{-1}[\mu(t)](x) \geq F(x), \quad x \geq 0,$$

and the new approximation is better than the preceding one at any point in $\mathbb{R}_+$.

The aim of step 2 is to reduce the error of a prior approximation, hence it can be carried out several times. However, in our examples, we will show that only one iteration of step 2 may be sufficient to obtain a reasonably small error.

Because μ must satisfy $\mu(0) = \lim_{t\to 0} \mu(t) = 0$, good candidates for μ are differences of Laplace transforms of stochastically ordered distributions. Taking, for instance, gamma or $\frac{1}{2}$-stable laws yields the forms

$$\mu(t) = c\left(\frac{a^\nu}{(a+t)^\nu} - \frac{b^\nu}{(b+t)^\nu}\right), \quad c, \nu > 0, \quad a > b > 0 \tag{3-3}$$

and

$$\mu(t) = c\big(e^{-\sqrt{at}} - e^{-\sqrt{bt}}\big), \quad c > 0, \quad b > a > 0. \tag{3-4}$$

We underline that the choice of a proper μ is crucial as it will enhance the approximation in a very peculiar way. For instance, choosing (3-3) will have a considerable impact on the tail of the Laplace approximation and thus on the behavior of the new CDF near 0; however, on the contrary, μ defined in (3-4) is negligible near infinity, but not near zero, thereby having the opposite effect on the target CDF: little impact near zero, but a significant modification of the tail of the approximating distribution.

Once μ is chosen (this task usually requires a fitting tool from a quantitative software), the critical point is to check that $h(t)/t := (N(t) - \mu(t))/t$ is indeed a completely monotone function. We recall that, for any C^n function h,

$$\begin{aligned}\left(\frac{h(\cdot)}{\cdot}\right)^{(n)}(t) &= \frac{1}{t^{n+1}} \sum_{i=0}^{n} \frac{(-1)^i n!}{(n-i)!} t^{n-i} h^{(n-i)}(t) \\ &= \frac{h^{(n)}(t) - n\,(h(\cdot)/\cdot)^{(n-1)}(t)}{t},\end{aligned} \tag{3-5}$$

which can be proven iteratively.

The function $h(t) - th'(t)$ requires a particular focus since it is associated with the first derivative. If the functions $t^n h^{(n)}(t)$ are smooth then some patterns can be identified for n small enough. If $h(\cdot)/\cdot$ is indeed completely monotone, then, in (3-5), the relative weight of $h^{(n)}$ compared to that of $n(h(\cdot)/\cdot)^{(n-1)}$ will decrease as n increases. The idea, based on empirical results, is to test to what extent

$$d_n(t) := -t\frac{(h(\cdot)/\cdot)^{(n)}(t)}{(h(\cdot)/\cdot)^{(n-1)}(t)} = n - \frac{h^{(n)}(t)}{(h(\cdot)/\cdot)^{(n-1)}(t)} \approx n \quad \text{for all } t \geq 0.$$

We provide examples below to illustrate this matter (Figures 1 and 2).

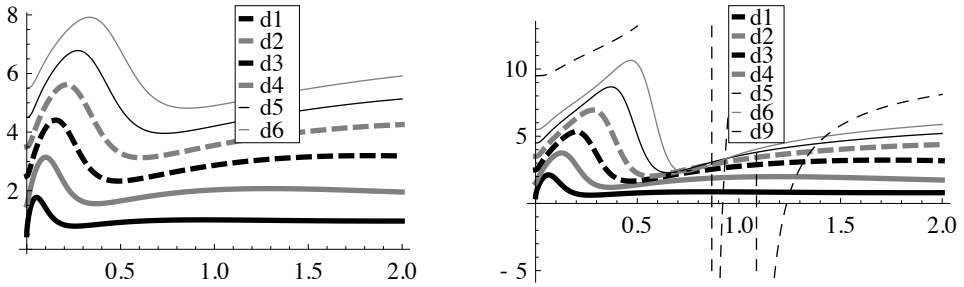


Figure 1. Graph of d_n for various n in two cases, for Example 4.1.

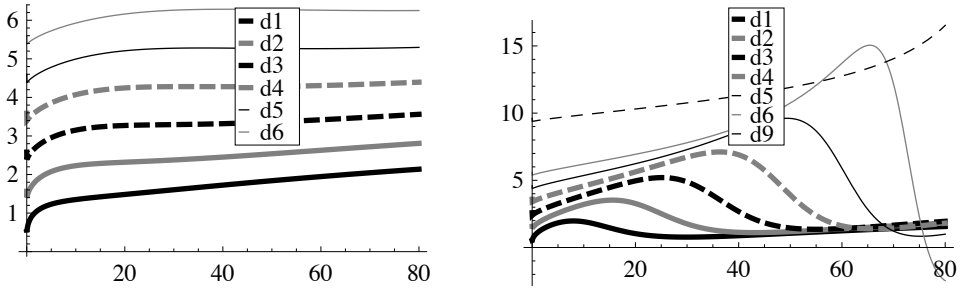


Figure 2. Graph of d_n for various n in two cases, for Example 4.2.

Among these four graphs, the two on the right fail the test: not only does d_n drift away from n, but there is a sign change at some point. The wave shape in three of the graphs is due to the fact that the function $h(t) - th'(t)$ has a local minimum away from zero. In this case, the successive derivatives may progressively (as n increases) hit zero in the vicinity of this local minimum. When there is no local minimum away from zero, our empirical tests have shown that the d_n are close to a constant or a slightly increasing affine function (as in the left graph of Figure 2).

Error results. We define $N = M - L$, $H = G - F$ and recall that $\mathbb{L}[H(x)](t) = N(t)/t$ is the Laplace transform of the error on the CDFs. The following proposition provides the Mellin transform of H (given N) and the Kantorovich distance between X and Y, which we define by

$$K(X, Y) = \sup\left\{\int_0^\infty f(x)(F(dx) - G(dx));\ f \in \mathrm{Lip}\right\}, \tag{3-6}$$

where Lip is the set of 1-Lipschitz functions. Dall'Aglio proved in [6] that, in fact,

$$K(X, Y) = \int_0^1 |F^{-1}(x) - G^{-1}(x)|\, dx = \int_0^\infty |F(x) - G(x)|\, dx$$

because the support of X and Y is $\mathbb{R}_+$.

We recall that in our setting, the functions N and H are either nonnegative or nonpositive. For simplicity, and without any loss of generality, we henceforth assume that they are nonnegative.

Proposition 3.2. *For $0 < b < 1$, whenever these integrals make sense,*

$$\int_0^\infty \frac{N(t)}{t^{1+b}} dt = \Gamma(1-b) \int_0^\infty x^{b-1} H(x)\, dx = \frac{\Gamma(1-b)}{b} \int_0^\infty H(x^{1/b})\, dx$$

Moreover,

$$\lim_{t \downarrow 0} N(t)/t = \int_0^\infty H(x)\, dx = K(X, Y) \tag{3-7}$$

Proof. The first equality is simply Fubini's theorem combined with the identity $\int_0^\infty e^{-xt} t^{-b} dt = \Gamma(1-b) x^{b-1}$ and a standard change of variable; the second equality is obvious. □

In some cases, it is possible to obtain an upper bound for the L^p quasi-norm of H for $p \in (0, 1)$, using Jensen's (reversed) inequality.

Lastly, we would like to recall the link between the Kantorovich distance and the Kolmogorov–Smirnov (uniform) distance $\sup_{x \geq 0} |F(x) - G(x)|$. Intercalating the Lévy and Prohorov metrics (using the results from [11, pp. 35–36] and [21, p. 43]), we get

$$\sup_{x \geq 0} |F(x) - G(x)| \leq \left((1+c) \int_0^\infty |F(x) - G(x)|\, dx \right)^{1/2}$$

where c is the maximum value (over $\mathbb{R}_+$) of $f = F'$, the density of X.

4. Examples

We test our method on two heavy-tailed distributions for which a rather simple closed form for f or F is available. The driving criterion for our approximations will be to get a finite Kantorovich distance.

Example 4.1. ***A generalized Mittag-Leffler distribution.*** We follow the notations of [13]. Generalized Mittag-Leffler distributions are a two-parameter family of laws with Laplace transforms

$$L(t) = (1 + t^\alpha)^{-\beta}, \quad \beta > 0,\ 0 < \alpha \leq 1$$

and cumulative distribution function

$$F(x) = \sum_{k=0}^{\infty} \frac{(-1)^k \Gamma(\beta + k) x^{\alpha(\beta+k)}}{k!\, \Gamma(\beta) \Gamma(1 + \alpha(\beta + k))}$$

We will focus on the simple case $\alpha = 1/2$, and $\beta = 2$. First notice that since $\frac{\Gamma(2k+2)}{\Gamma(k+2)(2k)!} = \frac{1}{k!}(2 - 1/(k+1))$,

$$\sum_{k=0}^{\infty} \frac{\Gamma(2+2k)x^{(2+2k)/2}}{(2k)!\,\Gamma(1+(2+2k)/2)} = e^x(2x-1)+1$$

Next, the odd integers are dealt with using the infinite series representation of the error function (8.253-1 in [9])

$$e^x \mathrm{erf}(\sqrt{x}) = \frac{2}{\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{2^k x^{k+1/2}}{(2k+1)!!}$$

where $(2k+1)!! = 1 \cdot 3 \cdot 5 \ldots (2k+1)$, and the identity

$$\frac{\Gamma(2k+3)}{(2k+1)!\,\Gamma(k+5/2)} = (2k+2)\frac{2^{k+2}}{\sqrt{\pi}(2k+3)!!}$$

which yields in the end

$$F(x) = e^x(2x-1)\mathrm{erfc}(\sqrt{x}) - 2\sqrt{x/\pi} + 1$$

From $L(t) = (1+\sqrt{t})^{-2}$, we know that $f(0+) = 1$ and that f has a heavy tail. We thus choose the Pareto family with $a = n$ in order to have the proper asymptotic behavior for L (when $t \to \infty$). In fact, the domination condition imposes $a, n \geq 1/2$ and a few tests show that $a = n = 1/2$ is a relevant choice, yielding

$$M(t) = \frac{\sqrt{t}e^{t/2}\Gamma(-1/2, t/2)}{2\sqrt{2}}, \quad t \geq 0$$

As expected, the approximation is not satisfactory and we must resort to an appropriate μ.

We wish to stress the importance of the choice of μ and we will test the performance of two functions, namely μ_1 and μ_2. The first naive choice was to take μ of the form

$$\mu_1(t) = c\left(\frac{a^{3/2}}{(a+t)^{3/2}} - \frac{b^{3/2}}{(b+t)^{3/2}}\right)$$

and an admissible set of parameters was $a = 3$, $b = 0.05$ and $c = 0.135$. This triple was the result of a fitting algorithm from a quantitative software.

Unfortunately, this approximation does not allow to compute the Kantorovich error because it is not good enough near 0. In order to be able to compute (3-7), we recall the expansion of L at zero (derived from that of $(1+t)^{-2}$):

$$(1+\sqrt{t})^{-2} = 1 - 2\sqrt{t} + 3t - 4t^{3/2} + O(t^2), \quad t \downarrow 0$$

Therefore, a strong improvement of the approximation should satisfy

$$M(t) - \mu_2(t) \sim 1 - 2\sqrt{t} + O(t), \quad t \downarrow 0.$$

By [20, 45:5:2] combined with [1, 6.5.17 and 7.1.5], we have

$$\frac{\sqrt{t}e^{t/2}\Gamma(-1/2, t/2)}{2\sqrt{2}} = 1 - \sqrt{\frac{\pi t e^t}{2}}\big(1 - \mathrm{erf}(\sqrt{t/2})\big) = 1 - \sqrt{\pi t/2} + t + O(t^{3/2})$$

as $t \downarrow 0$. Moreover,

$$e^{-\sqrt{at}} = 1 - \sqrt{at} + \tfrac{1}{2}at + O(t^{3/2}), \quad t \downarrow 0;$$

hence we propose $\mu_2(t) = c(e^{-\sqrt{at}} - e^{-\sqrt{bt}})$ with a, b, c satisfying $c(\sqrt{a} - \sqrt{b}) = -2 + \sqrt{\pi/2}$. The triple $a = 0.777$, $b = 20$ and $c = 0.206$ yields promising results with a Kantorovich distance of approximately 0.02 (computed via (3-7)).

Of course, in both cases, we have checked, using the d_n for $n \in \{1, \dots, 9\}$, that the error h was such that $h(t)/t$ was a completely monotone function.

We provide the graphical results below (Figures 3 and 4). M is the Laplace transform of the Pareto distribution with $a = n = 1/2$, $M_1(t) = M(t) - \mu_1(t)$ and $M_2(t) = M(t) - \mu_2(t)$. Their CDF counterparts are G, G_1 and G_2. It is plain on the graphs that M_1 and M_2 are quite close, except near zero; this explains why only G_2 is a good fit for F for x large (as expected).

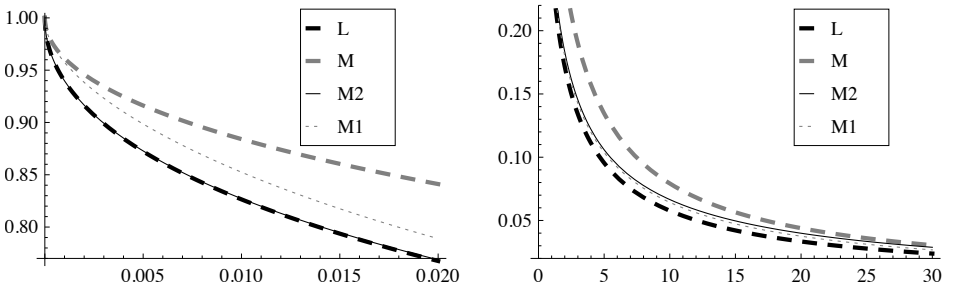


Figure 3. Graph of L and its proxies for $t \in (0, 0.02)$ and $t \in [0.02, 30]$.

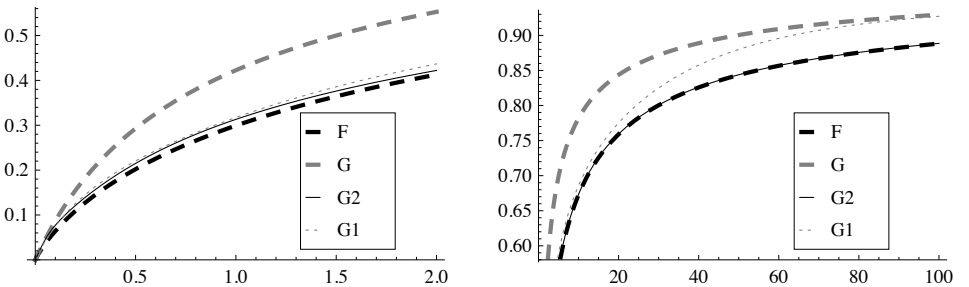


Figure 4. Graph of F and its proxies for $x \in (0, 2)$ and $x \in (2, 100)$.

Example 4.2. ***A positive stable distribution.*** Our next example is the one-parameter one-sided stable laws with Laplace transform

$$L(t) = e^{-t^\alpha}, \quad \alpha \in (0, 1)$$

The case $\alpha = \frac{1}{2}$ is sometimes referred to as the Lévy distribution, which is connected with the first passage time of the Brownian motion over fixed levels. The case $\alpha = \frac{1}{3}$ also has a closed-form density (see B.25 in [17] for instance):

$$f(x) = \frac{K_{1/3}\left(\sqrt{4/(27x)}\right)}{3\pi x^{3/2}}, \quad x \geq 0$$

We will thus aim at approximating $L(t) = e^{-t^{1/3}}$. In this case, $f(0+) = 0$ and f has a fat tail. Moreover,

$$L(t) = 1 - t^{1/3} + \tfrac{1}{2}t^{2/3} - \tfrac{1}{6}t + O(t^{4/3}), \quad t \downarrow 0 \tag{4-1}$$

The choice of the inverse gamma family with $n = 1/3$ seems relevant, as it satisfies (see 51:6:1 in [20])

$$M(t) = \frac{2}{\Gamma\left(\frac{1}{3}\right)}(at)^{1/6}K_{1/3}(2\sqrt{at}) = 1 + \frac{a^{1/3}\Gamma\left(-\frac{1}{3}\right)}{\Gamma\left(\frac{1}{3}\right)}t^{1/3} + 3at/2 + O(t^{4/3}), \quad t \downarrow 0.$$

Hence, for $a = -\Gamma\left(\frac{1}{3}\right)^3/\,\Gamma\left(-\frac{1}{3}\right)^3$, the $t^{1/3}$ term in the error will vanish, by (4-1), but the $t^{2/3}$ term will remain. This leads to the following choice of μ:

$$\begin{aligned}\mu(t) &= 2c\left(\frac{(bt)^{1/3}K_{2/3}(2\sqrt{bt})}{\Gamma\left(\frac{2}{3}\right)} - \frac{(dt)^{1/3}K_{2/3}(2\sqrt{dt})}{\Gamma\left(\frac{2}{3}\right)}\right)\\ &= c\frac{\Gamma\left(-\frac{2}{3}\right)}{\Gamma\left(\frac{2}{3}\right)}(b^{2/3} - d^{2/3})t^{2/3} + 3c(b-d)t + O(t^{5/3}), \quad t \downarrow 0.\end{aligned}$$

Notice that this time, the ordering is in the opposite way: $L(t) \geq M(t)$ for all $t \geq 0$. In this setting, an admissible set of parameters is $c = 6$, $b = 0.4$ and $d = 0.43$, which yields a Kantorovich distance of less than 0.06 (see Figures 5 and 6 on the next page, where $M_1(t) = M(t) + \mu(t)$).

Of course, in both examples, it is possible to further reduce the error by repeating step 2 on page 238 at least one time (using a minorant μ^* of $M - \mu - L$ for instance).

Remarks. We did not study distributions with lighter tails in the examples because when $1 - F(x) \leq cx^{-\alpha}$, with $c > 0$ and $\alpha > 1$ for any large x, then it is much easier to obtain a finite Kantorovich error, as the original survival function is already integrable.

Using the exact same procedure as in the second example, it would thus take $n - 2$ iterations of step 2 to obtain an approximation with finite Kantorovich error

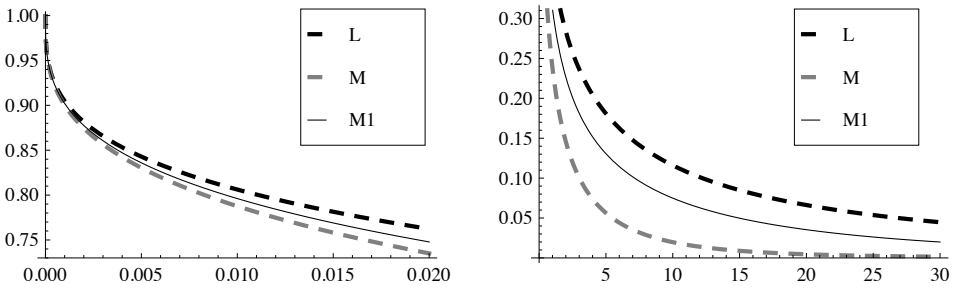


Figure 5. Graph of L and its proxies for $t \in (0, 0.02)$ and $t \in [0.02, 30]$.

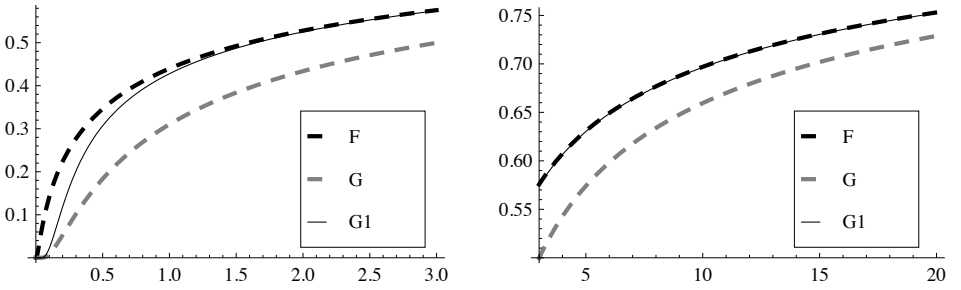


Figure 6. Graph of F and its proxies for $x \in (0, 2)$ and $x \in (2, 100)$.

for the stable law with Laplace transform equal to $e^{-t^{1/n}}$. The same holds for generalized Mittag-Leffler distributions defined by $L(t) = (1 + t^{1/n})^{-p}$, for any real $p \geq n$. These assertions are a consequence of the Taylor expansion of L at 0. In the stable case, when $\alpha \in \left(\frac{1}{2}, 1\right)$, it is possible to obtain a finite Kantorovich measure by taking $M(t) = e^{-\sqrt{t}}$ and μ such that $\mu(t) \sim t^{\alpha} - \sqrt{t}$ when $t \downarrow 0$.

Finally, we stress that even though we have assumed (for simplicity) that the law of X was absolutely continuous, our method remains valid for most positive laws. It is indeed possible to make do without densities throughout the whole process, especially if the law of X has a finite number of atoms and an absolutely continuous part. However, it is not clear whether this method can perform well for some rather unusual distributions, such as those which possess an infinite number of atoms.

References

[1] M. Abramowitz and I. A. Stegun (eds.), *Handbook of mathematical functions, with formulas, graphs and mathematical tables*, National Bureau of Standards Applied Mathematics Series, no. 55, National Bureau of Standards, Washington, DC, 1966. MR 34 #8607 Zbl 0643.33001

[2] A. Al-Shuaibi, *Inversion of the Laplace transform via Post–Widder formula*, Integral Transform. Spec. Funct. **11** (2001), no. 3, 225–232. MR 2003d:44001 Zbl 1022.65138

[3] N. H. Bingham, C. M. Goldie, and J. L. Teugels, *Regular variation*, Encyclopedia of Mathematics and its Applications, no. 27, Cambridge University Press, Cambridge, 1987. MR 88i:26004 Zbl 0617.26001

[4] K. F. Chen and S. L. Mei, *Accelerations of Zhao's methods for the numerical inversion of Laplace transform*, Int. J. Numer. Methods Biomed. Eng. **27** (2011), no. 2, 273–282. MR 2011k:65185 Zbl 1211.65167

[5] A. M. Cohen, *Numerical methods for Laplace transform inversion*, Numerical Methods and Algorithms, no. 5, Springer, New York, 2007. MR 2009c:65354 Zbl 1127.65094

[6] G. Dall'Aglio, *Sugli estremi dei momenti delle funzioni di ripartizione doppia*, Ann. Scuoloa Norm. Sup. Pisa (3) **10** (1956), 35–74. MR 18,423i Zbl 0073.14002

[7] C. L. Epstein and J. Schotland, *The bad truth about Laplace's transform*, SIAM Rev. **50** (2008), no. 3, 504–520. MR 2009g:44002 Zbl 1154.65094

[8] W. Feller, *An introduction to probability theory and its applications*, 2nd ed., vol. II, Wiley, New York, 1971. MR 42 #5292 Zbl 0219.60003

[9] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, 7th ed., Elsevier, Amsterdam, 2007. MR 2008g:00005 Zbl 1208.65001

[10] C.-Y. Hu and G. D. Lin, *Some inequalities for Laplace transforms*, J. Math. Anal. Appl. **340** (2008), no. 1, 675–686. MR 2009e:44001 Zbl 1156.60012

[11] P. J. Huber and E. M. Ronchetti, *Robust statistics*, 2nd ed., Wiley, Hoboken, NJ, 2009. MR 2010j:62004 Zbl 00520286

[12] P. Jara, F. Neubrander, and K. Özer, *Rational inversion of the Laplace transform*, J. Evol. Equ. **12** (2012), no. 2, 435–457. MR 2923942 Zbl 06113139

[13] K. K. Jose, P. Uma, V. S. Lekshmi, and H. J. Haubold, *Generalized Mittag-Leffler distributions and processes for applications in astrophysics and time series modeling*, Proceedings of the Third UN/ESA/NASA Workshop on the International Heliophysical Year 2007 and Basic Space Science (Tokyo, 2007) (H. J. Haubold and A. M. Mathai, eds.), Springer, Heidelberg, 2010, pp. 79–92. MR 2011g:62242 Zbl 1223.85011

[14] F.-R. Lin and F. Liang, *Application of high order numerical quadratures to numerical inversion of the Laplace transform*, Adv. Comput. Math. **36** (2012), no. 2, 267–278. MR 2886191 Zbl 06038160

[15] I. M. Longman, *Numerical Laplace transform inversion of a function arising in viscoelasticity*, J. Comput. Phys. **10** (1972), no. 2, 224–231 (English). Zbl 0246.65034

[16] Y. L. Luke, *Error estimation in numerical inversion of Laplace transforms using Padé approximation*, Journal of the Franklin Institute **305** (1978), no. 5, 259–273.

[17] F. Mainardi, P. Paradisi, and R. Gorenflo, *Probability distributions generated by fractional diffusion equations*, Workshop on Econophysics (Budapest, 1997), 2007. arXiv 0704.0320

[18] V. Masol and J. L. Teugels, *Numerical accuracy of real inversion formulas for the Laplace transform*, J. Comput. Appl. Math. **233** (2010), no. 10, 2521–2533. MR 2011a:65447 Zbl 1183.65172

[19] K. Nakagawa, *Tail probability and singularity of Laplace–Stieltjes transform of a heavy tailed random variable*, Information Theory and Its Applications (ISITA) (Auckland, 2008), IEEE, Piscataway, NJ, 2009. arXiv 0909.0090

[20] K. Oldham, J. Myland, and J. Spanier, *An atlas of functions: with Equator, the atlas function calculator*, 2nd ed., Springer, New York, 2009. MR 2010f:33001 Zbl 1167.65001

[21] V. V. Petrov, *Limit theorems of probability theory: sequences of independent random variables*, Oxford Studies in Probability, no. 4, Clarendon/Oxford University Press, New York, 1995. MR 96h:60048 Zbl 0826.60001

[22] A. D. Polyanin and A. V. Manzhirov, *Handbook of integral equations*, 2nd ed., CRC, Boca Raton, FL, 2008. MR 2009a:45001 Zbl 1154.45001

[23] A. Stef and G. Tenenbaum, *Inversion de Laplace effective*, Ann. Probab. **29** (2001), no. 1, 558–575. MR 2002c:60026 Zbl 1020.60008

[24] V. K. Tuan and D. T. Duc, *Convergence rate of Post–Widder approximate inversion of the Laplace transform*, Vietnam J. Math. **28** (2000), no. 1, 93–96. MR 1811305 Zbl 0969.44001

[25] M. S. Veillette and M. S. Taqqu, *A technique for computing the PDFs and CDFs of non-negative infinitely divisible random variables*, J. Appl. Probab. **48** (2011), no. 1, 217–237. MR 2012d:60045 Zbl 1210.60023

[26] J. A. C. Weideman, *Optimizing Talbot's contours for the inversion of the Laplace transform*, SIAM J. Numer. Anal. **44** (2006), no. 6, 2342–2362. MR 2007k:65219 Zbl 1131.65105

[27] H. Wendland, *Scattered data approximation*, Cambridge Monographs on Applied and Computational Mathematics, no. 17, Cambridge University Press, Cambridge, 2005. MR 2006i:41002 Zbl 1075.65021

Received March 23, 2012. Revised July 27, 2012.

GUILLAUME COQUERET: guillaume.coqueret@essec.edu
ESSEC Business School / Université de Lille-1, Avenue Bernard Hirsch, 95000 Cergy-Pontoise, France

COMM. APP. MATH. AND COMP. SCI.
Vol. 7, No. 2, 2012
dx.doi.org/10.2140/camcos.2012.7.247

msp

OPTIMAL STABILITY POLYNOMIALS FOR NUMERICAL INTEGRATION OF INITIAL VALUE PROBLEMS

DAVID I. KETCHESON AND ARON J. AHMADIA

We consider the problem of finding optimally stable polynomial approximations to the exponential for application to one-step integration of initial value ordinary and partial differential equations. The objective is to find the largest stable step size and corresponding method for a given problem when the spectrum of the initial value problem is known. The problem is expressed in terms of a general least deviation feasibility problem. Its solution is obtained by a new fast, accurate, and robust algorithm based on convex optimization techniques. Global convergence of the algorithm is proven in the case that the order of approximation is one and in the case that the spectrum encloses a starlike region. Examples demonstrate the effectiveness of the proposed algorithm even when these conditions are not satisfied.

1. Stability of Runge–Kutta methods

Runge–Kutta methods are among the most widely used types of numerical integrators for solving initial value ordinary and partial differential equations. The time step size should be taken as large as possible since the cost of solving an initial value problem (IVP) up to a fixed final time is proportional to the number of steps that must be taken. In practical computation, the time step is often limited by stability and accuracy constraints. Either accuracy, stability, or both may be limiting factors for a given problem; see [24, Section 7.5] for a discussion. The linear stability and accuracy of an explicit Runge–Kutta method are characterized completely by the so-called stability polynomial of the method, which in turn dictates the acceptable step size [6; 12]. In this work we present an approach for constructing a stability polynomial that allows the largest absolutely stable step size for a given problem.

The problem of finding optimal stability polynomials is of fundamental importance in the numerical solution of initial value problems, and its solution or approximation has been studied by many authors for several decades Indeed, it is closely related to the problem of finding polynomials of least deviation, which goes back to the work of Chebyshev. A nice review of much of the early work on

MSC2010: primary 65L06, 65M20; secondary 90C26.
Keywords: absolute stability, initial value problems, Runge–Kutta methods.

Runge–Kutta stability regions can be found in [44]. The most-studied cases are those where the eigenvalues lie on the negative real axis [1; 3; 2; 4; 38; 23; 25; 27; 33; 35; 36; 43; 8], on the imaginary axis [21; 20; 22; 26; 43; 32; 46], or in a disk of the form $|z + w| \leq w$ [15; 46]. Many results and optimal polynomials, both exact and numerical, are available for these cases. Some authors have considered the solution of Problem 1 for other spectra corresponding to PDE semidiscretizations [17; 31; 38; 26; 28; 39].

Two very recent works serve to illustrate both the progress that has been made in solving these problems with nonlinear programming, and the challenges that remain. In [39], optimal schemes are sought for integration of discontinuous Galerkin discretizations of wave equations, where the optimality criteria considered include both accuracy and stability measures. The approach used there is based on sequential quadratic programming (local optimization) with many initial guesses. The authors consider methods of at most fourth order and situations with $s - p \leq 4$ "because the cost of the optimization procedure becomes prohibitive for a higher number of free parameters." In [28], optimally stable polynomials are found for certain spectra of interest for $2 \leq p \leq 4$ and (in a remarkable feat!) s as large as 14. The new methods obtained achieve a 40–50% improvement in efficiency for discontinuous Galerkin integration of the 3D Maxwell equations. The optimization approach employed therein is again a direct search algorithm that does not guarantee a globally optimal solution but "typically converges... within a few minutes". However, it was apparently unable to find solutions for $s > 14$ or $p > 4$. The method we present in the next section can rapidly find solutions for significantly larger values of s, p, and is provably globally convergent under certain assumptions (introduced in Section 2).

In the remainder of this section, we review the stability concepts for Runge–Kutta methods and formulate the stability optimization problem. Our optimization approach, described in Section 2, is based on reformulating the stability optimization problem in terms of a sequence of convex subproblems and using bisection. We examine the theoretical properties of the proposed algorithm and prove its global convergence for two important cases.

A key element of our optimization algorithm is the use of numerical convex optimization techniques. We avoid a poorly conditioned numerical formulation by posing the problem in terms of a polynomial basis that is well-conditioned when sampled over a particular region of the complex plane. These numerical considerations, which become particularly important when the number of stages of the method is allowed to be very large, are discussed in Section 3.

In Section 4 we apply our algorithm to several examples of complex spectra. Cases where optimal results are known provide verification of the algorithm, and many new or improved results are provided.

Determination of the stability polynomial is only half of the puzzle of designing optimal explicit Runge–Kutta methods. The other half is the determination of the Butcher coefficients. While simply finding methods with a desired stability polynomial is straightforward, many additional challenges arise in that context; for instance, additional nonlinear order conditions, internal stability, storage, and embedded error estimators. All of these concerns can be dealt with using the software package RK-opt [19], which also includes the algorithm presented herein. The development of full Runge–Kutta methods based on optimal stability polynomials, using the present approach and additional tools from RK-opt, is conducted in [30].

1.1. ***The stability polynomial.*** A linear, constant-coefficient initial value problem takes the form

$$u'(t) = Lu, \qquad u(0) = u_0, \tag{1}$$

where $u(t) : \mathbb{R} \to \mathbb{R}^N$ and $L \in \mathbb{R}^{N\times N}$. When applied to the linear IVP (1), any Runge–Kutta method reduces to an iteration of the form

$$u_n = R(hL)u_{n-1}, \tag{2}$$

where h is the step size and u_n is a numerical approximation to $u(nh)$. The *stability function* $R(z)$ depends only on the coefficients of the Runge–Kutta method; see [9, Section 4.3], [6], [12]. In general, the stability function of an s-stage explicit Runge–Kutta method is a polynomial of degree s

$$R(z) = \sum_{j=0}^{s} a_j z^j. \tag{3}$$

Recall that the exact solution of (1) is $u(t) = \exp(tL)u_0$. Thus, if the method is accurate to order p, the stability polynomial must be identical to the exponential function up to terms of at least order p:

$$a_j = \frac{1}{j!} \quad \text{for } 0 \le j \le p. \tag{4}$$

1.2. ***Absolute stability.*** The stability polynomial governs the local propagation of errors, since any perturbation to the solution will be multiplied by $R(z)$ at each subsequent step. The propagation of errors thus depends on $\|R(hL)\|$, which leads us to define the *absolute stability region*

$$S = \{z \in \mathbb{C} : |R(z)| \le 1\}. \tag{5}$$

For example, the stability region of the classical fourth-order method is shown in Figure 1(b).

given an initial value problem (1), let $\Lambda \in \mathbb{C}$ denote the spectrum of the matrix L. We say the iteration (2) is absolutely stable if

$$h\lambda \in S \quad \text{for all } \lambda \in \Lambda. \tag{6}$$

Condition (6) implies that u_n remains bounded for all n. More importantly, (6) is a necessary condition for stable propagation of errors. Thus the maximum stable step size is given by

$$h_{\text{stable}} = \max\{h \geq 0 : |R(h\lambda)| \leq 1 \text{ for } \lambda \in \Lambda\}. \tag{7}$$

Note that for nonnormal L, it may be important to consider the pseudospectrum rather than the spectrum; see Section 4.3.

As an example, consider the advection equation

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}u(x,t) = 0, \quad x \in (0, M),$$

discretized in space by first-order upwind differencing with spatial mesh size Δx

$$U_i'(t) = -\frac{U_i(t) - U_{i-1}(t)}{\Delta x}, \quad 0 \leq i \leq N,$$

with periodic boundary condition $U_0(t) = U_N(t)$. This is a linear IVP (1) with L a circulant bidiagonal matrix. The eigenvalues of L are plotted in Figure 1(a) for $\Delta x = 1$, $N = M = 20$. To integrate this system with the classical fourth-order Runge–Kutta method, the time step size must be taken small enough that the scaled spectrum $\{h\lambda_i\}$ lies inside the stability region. Figure 1(c) shows the (maximally) scaled spectrum superimposed on the stability region.

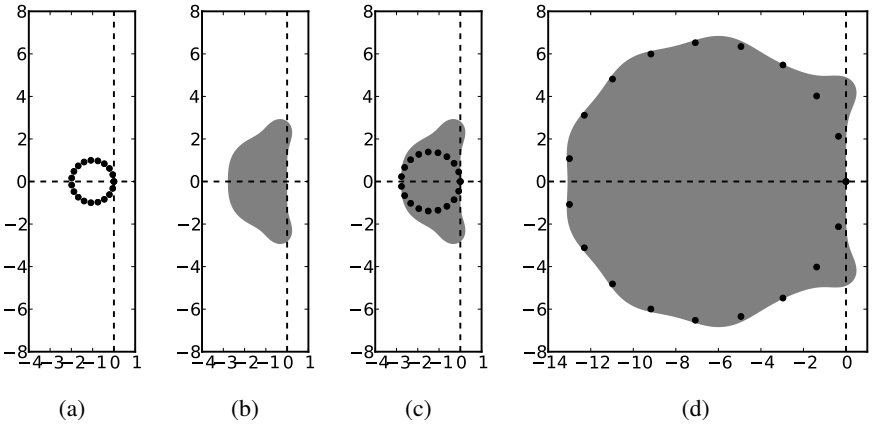


Figure 1. (a) Spectrum of first-order upwind difference matrix using $N = 20$ points in space. (b) Stability region of the classical fourth-order Runge–Kutta method. (c) Scaled spectrum $h\lambda$ with $h = 1.39$. (d) Scaled spectrum $h\lambda$ for optimal ten-stage method with $h = 6.54$.

The motivation for this work is that a larger stable step size can be obtained by using a Runge–Kutta method with a larger region of absolute stability. Figure 1(d) shows the stability region of an optimized ten-stage Runge–Kutta method of order four that allows a much larger step size. The ten-stage method was obtained using the technique that is the focus of this work. Since the cost of taking one step is typically proportional to the number of stages s, we can compare the efficiency of methods with different numbers of stages by considering the *effective step size* h/s. Normalizing in this manner, it turns out that the ten-stage method is nearly twice as fast as the traditional four-stage method.

1.3. ***Design of optimal stability polynomials.*** We now consider the problem of choosing a stability polynomial so as to maximize the step size under which given stability constraints are satisfied. The objective function $f(x)$ is simply the step size h. The stability conditions yield nonlinear inequality constraints. Typically one also wishes to impose a minimal order of accuracy. The monomial basis representation (3) of $R(z)$ is then convenient because the first $p+1$ coefficients $\{a_0, a_1, \ldots, a_p\}$ of the stability polynomial are simply taken to satisfy the order conditions (4). As a result, the space of decision variables has dimension $s+1-p$, and is comprised of the coefficients $\{a_{p+1}, a_{p+2}, \ldots, a_s\}$, as well as the step size h. Then the problem can be written as

Problem 1 (stability optimization). Given $\Lambda \subset \mathbb{C}$, order p, and number of stages s,

$$\begin{aligned} &\underset{a_{p+1},a_{p+2},\ldots,a_s,h}{\text{maximize}} && h \\ &\text{subject to} && |R(h\lambda)| - 1 \le 0 \quad \text{for all } \lambda \in \Lambda. \end{aligned}$$

We use H_{opt} to denote the solution of Problem 1 (the optimal step size) and R_{opt} to denote the optimal polynomial.

The set Λ may be finite, corresponding to a finite-dimensional ODE system or PDE semidiscretization, or infinite (but bounded), corresponding to a PDE or perhaps its semidiscretization in the limit of infinitesimal mesh width. In the latter case, Problem 1 is a semi-infinite program (SIP). In Section 4 we approach this by using a finite discretization of Λ; for a discussion of this and other approaches to semi-infinite programming, see [13].

2. An efficient algorithm for design of globally optimal stability polynomials

Evidently, finding the global solution of Problem 1 is in general quite challenging. Although sophisticated optimization algorithms such as the interior point method can guarantee polynomial time solutions to convex problems, and convex programming techniques are valuable in efficiently seeking minima, the stability constraints in Problem 1 are nonconvex. As a result, suboptimal local minima may exist and

certificates of optimality may require either approximations to the solution of the problem or greater than polynomial time. See [5] for an overview of convex optimization programming techniques, and [29] for an introduction to approximation algorithms and local search heuristics for nonconvex problems.

2.1. ***Reformulation in terms of the least deviation problem.*** The primary theoretical advance leading to the new results in this paper is a reformulation of Problem 1. Note that Problem 1 is nonconvex for $s > 2$ since $R(h\lambda)$ is a nonconvex function in h.

Instead of asking for the maximum stable step size we now ask, for a given step size h, how small the maximum modulus of $R(h\lambda)$ can be. This leads to a generalization of the classical least deviation problem.

Problem 2 (least deviation). Given $\Lambda \subset \mathbb{C}$, $h \in \mathbb{R}^+$ and $p, s \in \mathbb{N}$

$$\underset{a_{p+1},a_{p+2},\ldots,a_s}{\text{minimize}} \quad \max_{\lambda\in\Lambda}\big(|R(h\lambda| - 1\big).$$

We denote the solution of Problem 2 by $r_{p,s}(h, \Lambda)$, or simply $r(h, \Lambda)$. Note that $|R(z)|$ is convex with respect to a_j, since $R(z)$ is linear in the a_j. Therefore, Problem 2 is convex. Problem 1 can be formulated in terms of Problem 2:

Problem 3 (reformulation of Problem 1). Given $\Lambda \subset \mathbb{C}$, and $p, s \in \mathbb{N}$,

$$\begin{aligned} &\underset{a_{p+1},a_{p+2},\ldots,a_s}{\text{maximize}} \quad h \\ &\text{subject to} \quad r_{p,s}(h, \Lambda) \le 0. \end{aligned}$$

2.2. ***Solution via bisection.*** Although Problem 3 is not known to be convex, it is an optimization in a single variable. It is natural then to apply a bisection approach, as outlined in Algorithm 1.

Algorithm 1 (simple bisection).

```
Select h_max (see Section 2.3)
h_min = 0
while h_max − h_min > ϵ do
    h = (h_max + h_min)/2
    Solve Problem 2
    if r_{p,s}(h, Λ) ≤ 0 then
        h_min = h
    else
        h_max = h
    end if
end while
return H_ϵ = h_min
```

Since $r(0, \Lambda) = -1$ and $\lim_{h\to\infty} r(h, \Lambda) = +\infty$, then there exists $h_{\max} > 0$ such that $r(h, \Lambda) = 0$ for some $h \in [h_{\min}, h_{\max}]$. Global convergence of the algorithm is assured only if the following condition holds:

$$r_{p,s}(h_0, \Lambda) = 0 \quad \Longrightarrow \quad r_{p,s}(h, \Lambda) \le 0 \ \text{ for all } \ 0 \le h \le h_0. \tag{8}$$

We now consider conditions under which condition (8) can be established. We have the following important case.

Theorem 1 (global convergence when $p = 1$). *Let $p = 1$, $\Lambda \subset \mathbb{C}$ and $s \ge 1$. Take $h_{\max}$ large enough so that $r(h_{\max}, \Lambda) > 0$. Let H_{opt} denote the solution of Problem 1. Then the output of Algorithm 1 satisfies*

$$\lim_{\epsilon\to 0} H_\epsilon = H_{\text{opt}}.$$

Proof. Since $r(0, \Lambda) = 0 < r(h_{\max}, \Lambda)$ and $r(h, \Lambda)$ is continuous in h, it is sufficient to prove that condition (8) holds. We have $|R_{\text{opt}}(H_{\text{opt}}\lambda)| \le 1$ for all $\lambda \in \Lambda$. We will show that there exists $R_\mu(z) = \sum_{j=0}^{s} a_j(\mu) z^j$ such that $a_0 = a_1 = 1$ and

$$|R_\mu(\mu H_{\text{opt}}\lambda)| \le 1 \quad \text{for all } \ \lambda \in \Lambda \text{ and } 0 \le \mu \le 1.$$

Let $\hat{a}_j$ be the coefficients of the optimal polynomial:

$$R_{\text{opt}}(z) = 1 + z + \sum_{j=2}^{s} \hat{a}_j z^j,$$

and set

$$a_j(\mu) = \mu^{1-j} \hat{a}_j.$$

Then

$$\begin{aligned} R_\mu(\mu H_{\text{opt}}\lambda) &= 1 + \mu H_{\text{opt}}\lambda + \sum_{j=2}^{s} \mu^{1-j} \hat{a}_j \mu^j H_{\text{opt}}^j \lambda^j = 1 + \mu \left(\sum_{j=1}^{s} \hat{a}_j H_{\text{opt}}^j \lambda^j \right) \\ &= 1 + \mu(R_{\text{opt}}(H_{\text{opt}}\lambda) - 1), \end{aligned}$$

where we have defined $\hat{a}_1 = 1$. Define $g_\lambda(\mu) = R_\mu(\mu H_{\text{opt}}\lambda)$. Then $g_\lambda(\mu)$ is linear in μ and has the property that, for $\lambda \in \Lambda$, $|g_\lambda(0)| = 1$ and $|g_\lambda(1)| \le 1$ (by the definition of H_{opt}, R_{opt}). Thus by convexity $|g(\mu)| \le 1$ for $0 \le \mu \le 1$. □

For $p > 1$, condition (8) does not necessarily hold. For example, take $s = p = 4$; then the stability polynomial (3) is uniquely defined as the degree-four Taylor approximation of the exponential, corresponding to the classical fourth-order Runge–Kutta method that we saw in the introduction. Its stability region is plotted in Figure 1(b). Taking, e.g., $\lambda = 0.21 + 2.3i$, one finds $|R(\lambda)| < 1$ but $|R(\lambda/2)| > 1$. Although this example shows that Algorithm 1 might formally fail, it concerns only the trivial case $s = p$ in which there is only one possible choice of stability

polynomial. We have searched without success for a situation with $s > p$ for which condition (8) is violated.

2.3. ***Selection of $h_{\max}$.*** The bisection algorithm requires as input an initial $h_{\max}$ such that $r(h_{\max}, \Lambda) > 0$. Theoretical values can be obtained using the classical upper bound of $2s^2/x$ if Λ encloses a negative real interval $[x, 0]$, or using the upper bound given in [34] if Λ encloses an ellipse in the left half-plane. Alternatively, one could start with a guess and successively double it until $r(h_{\max}, \Lambda) > 0$ is satisfied. Since evaluation of $r(h, \Lambda)$ is typically quite fast, finding a tight initial $h_{\max}$ is not an essential concern.

2.4. ***Convergence for starlike regions.*** In many important applications the relevant set Λ is an infinite set; for instance, if we wish to design a method for some PDE semidiscretization that will be stable for any spatial discretization size. In this case, Problem 1 is a semi-infinite program (SIP) as it involves infinitely many constraints. Furthermore, Λ is often a closed curve whose interior is starlike with respect to the origin; for example, upwind semidiscretizations of hyperbolic PDEs have this property. Recall that a region S is starlike if $t \in S$ implies $\mu t \in S$ for all $0 \le \mu \le 1$.

Lemma 1. *Let $\Lambda \in \mathbb{C}$ be a closed curve passing through the origin and enclosing a starlike region. Let $r(h, \Lambda)$ denote the solution of Problem 2. Then condition* (8) *holds.*

Proof. Let Λ be as stated in the lemma. Suppose $r(h_0, \Lambda) = 0$ for some $h_0 > 0$; then there exists $R(z)$ such that $|R(h\lambda)| \le 1$ for all $\lambda \in \Lambda$. According to the maximum principle, the stability region of $R(z)$ must contain the region enclosed by Λ. Choose h such that $0 \le h \le h_0$; then $h\Lambda$ lies in the region enclosed by Λ, so $|R(h\lambda)| \le 1$ for $\lambda \in \Lambda$. □

The proof of Lemma 1 relies crucially on Λ being an infinite set, but in practice we numerically solve Problem 2 with only finitely many constraints. To this end we introduce a sequence of discretizations Λ_n with the following properties:

1. $\Lambda_n \subset \Lambda$.
2. $n_1 \le n_2 \Longrightarrow \Lambda_{n_1} \subset \Lambda_{n_2}$.
3. $\lim_{n\to\infty} \Lambda_n = \Lambda$.
4. $\lim_{n\to\infty} \nu_n = 0$ where ν_n denotes the maximum distance from a point in Λ to the set Λ_n:

$$\nu_n = \max_{\gamma\in\Lambda} \min_{\lambda\in\Lambda_n} |\gamma - \lambda|.$$

For instance, Λ_n can be taken as an equispaced (in terms of arc-length, say) sampling of n points.

By modifying Algorithm 1, we can approximate the solution of the semi-infinite programming problem for starlike regions to arbitrary accuracy. At each step we solve Problem 2 with Λ_n replacing Λ. The key to the modified algorithm is to only increase $h_{\min}$ after obtaining a certificate of feasibility. This is done by using the Lipschitz constant of $R(z)$ over a domain including $h\Lambda$ (denoted by $L(R, h\Lambda)$) to ensure that $|R(h\Lambda)| \leq 1$. The modified algorithm is stated as Algorithm 2.

Algorithm 2 (bisection for SIP).

$h_{\min} = 0$
$h_{\max} = 2s^2/\max|\lambda|$
$n = n_0$
while $h_{\max} - h_{\min} > \epsilon$ **do**
  $h = (h_{\max} + h_{\min})/2$ ▷ Bisect
  Solve Problem 2
  if $r(h, \Lambda_n) < 0$ and $v_n < -2r/L(R, h\Lambda)$ **then** ▷ Certifies that $r(h, \Lambda) < 0$
    $h_{\min} = h$
  else if $r(h, \Lambda_n) > 0$ **then** ▷ Certifies that $r(h, \Lambda) > 0$
    $h_{\max} = h$
  else ▷ $r(h, \Lambda_n) \leq 0$
    $n \leftarrow 2n$ ▷ Reduce the discretization spacing
  end if
end while
return $H_\epsilon = h_{\min}$

The following lemma, which characterizes the behavior of Algorithm 2, holds whether or not the interior of Λ is starlike.

Lemma 2. *Let $h^{[k]}$ denote the value of h after k iterations of the loop in Algorithm 2. Then either*

- *Algorithm 2 terminates after a finite time with outputs satisfying $r(h_{\min}, \Lambda) \leq 0$, $r(h_{\max}, \Lambda) > 0$; or*
- *there exists $j < \infty$ such that $r(h^{[j]}, \Lambda) = 0$ and $h^{[k]} = h^{[j]}$ for all $j \geq k$.*

Proof. First suppose that $r(h^{[j]}, \Lambda) = 0$ for some j. Then neither feasibility nor infeasibility can be certified for this value of h, so $h^{[k]} = h^{[j]}$ for all $j \geq k$.

On the other hand, suppose that $r(h^{[k]}, \Lambda) \neq 0$ for all k. The algorithm will terminate as long as, for each $h^{[k]}$, either feasibility or infeasibility can be certified for large enough n. If $r(h^{[k]}, \Lambda) > 0$, then necessarily $r(h^{[k]}, \Lambda_n) > 0$ for large enough n, so infeasibility will be certified. We will show that if $r(h^{[k]}, \Lambda) < 0$, then for large enough n the condition

$$v_n < -2r/L(R, h\Lambda) \tag{9}$$

must be satisfied. Since $r(h, \Lambda_n) \leq r(h, \Lambda)$ is bounded away from zero and $\lim_{n\to\infty} \nu_n = 0$, (9) must be satisfied for large enough n unless the Lipschitz constant $L(R, h\Lambda)$ is unbounded (with with respect to n) for some fixed h. Suppose by way of contradiction that this is the case, and let $R^{[1]}, R^{[2]}, \ldots$ denote the corresponding sequence of optimal polynomials. Then the norm of the vector of coefficients $a_j^{[i]}$ appearing in $R^{[i]}$ must also grow without bound as $i \to \infty$. By Lemma 3, this implies that $|R^{[i]}(z)|$ is unbounded except for at most s points $z \in \mathbb{C}$. But this contradicts the condition $|R^{[i]}(h\lambda)| \leq 1$ for $\lambda \in \Lambda_n$ when $n > s$. Thus, for large enough n we must have $\nu_n < -2r/L(R, h\Lambda)$. □

In practical application, $r(h, \Lambda) = 0$ will not be detected, due to numerical errors; see Section 3.1. For this reason, in the next theorem we simply assume that Algorithm 2 terminates. We also require the following technical result.

Lemma 3. *Let $R^{[1]}, R^{[2]}, \ldots$ be a sequence of polynomials of degree at most s ($s \in \mathbb{N}$ fixed) and denote the coefficients of $R^{[i]}$ by $a_j^{[i]} \in \mathbb{C}$ ($i \in \mathbb{N}, 0 \leq j \leq s$):*

$$R^{[i]}(z) = \sum_{j=0}^{s} a_j^{[i]} z^j, \quad z \in \mathbb{C}.$$

Further, let $a^{[i]} := (a_0^{[i]}, a_1^{[i]}, \ldots, a_s^{[i]})^T$ and suppose that the sequence $\|a^{[i]}\|$ is unbounded in $\mathbb{R}$. Then the sequences $R^{[i]}(z)$ are unbounded for all but at most s points $z \in \mathbb{C}$.

Proof. Suppose to the contrary there are $s + 1$ *distinct* complex numbers, say, $z_0, z_1, \ldots, z_s$ such that the vectors $r_i := (R^{[i]}(z_0), R^{[i]}(z_1), \ldots, R^{[i]}(z_s))^T$ ($i \in \mathbb{N}$) are *bounded* in $\mathbb{C}^{s+1}$. Let V denote the $(s + 1) \times (s + 1)$ Vandermonde matrix whose k^{th} row ($0 \leq k \leq s + 1$) is $(1, z_k, z_k^2, \ldots, z_k^s)$. Then V is invertible and we have $a^{[i]} = V^{-1} r_i$ ($i \in \mathbb{N}$), so if $|||\cdot|||$ denotes the induced matrix norm, then

$$\|a^{[i]}\| = \|V^{-1} r_i\| \leq |||V^{-1}||| \, \|r_i\|.$$

But, by assumption, the right side is bounded, whereas the left side is not. □

Theorem 2 (global convergence for strictly starlike regions). *Let Λ be a closed curve that encloses a region that is starlike with respect to the origin. Suppose that Algorithm 2 terminates for all small enough ϵ, and let H_ϵ denote the value returned by Algorithm 2 for a given ϵ. Let H_{opt} denote the solution of Problem 1. Then*

$$\lim_{\epsilon \to 0} H_\epsilon = H_{\text{opt}}.$$

Proof. Due to the assumptions and Lemma 2, we have that $r(h_{\min}, \Lambda) < 0 < r(h_{\max}, \Lambda)$. Then Lemma 1 implies that $h_{\min} < H_{\text{opt}} < h_{\max}$. Noting that also $h_{\max} - h_{\min} < \epsilon$, the result follows. □

Despite the lack of a general global convergence proof, Algorithm 1 works very well in practice even for general Λ when $p > 1$. In all cases we have tested and for which the true H_{opt} is known (see Section 4), Algorithm 1 appears to converge to the globally optimal solution. Furthermore, Algorithm 1 is very fast. For these reasons, we consider the (much slower) Algorithm 2 to be of primarily theoretical interest, and we base our practical implementation on Algorithm 1.

3. Numerical implementation

We have made a prototype implementation of Algorithm 1 in Matlab. The implementation relies heavily on the CVX package [11; 10], a Matlab-based modeling system for convex optimization, which in turn relies on the interior-point solvers SeDuMi [37] and SDPT3 [42]. The least deviation problem (Problem 2) can be succinctly stated in four lines of the CVX problem language, and for many cases is solved in under a second by either of the core solvers.

Our implementation re-attempts failed solves (see Section 3.2) with the alternate interfaced solver. In our test cases, we observed that the SDPT3 interior-point solver was slower, but more robust than SeDuMi. Consequently, our prototype implementation uses SDPT3 by default.

Using the resulting implementation, we were able to successfully solve problems to within 0.1% accuracy or better with scaled eigenvalue magnitudes $|h\lambda|$ as large as 4000. As an example, comparing with results of [4] for spectra on the real axis with $p = 3, s = 27$, our results are accurate to 6 significant digits.

3.1. ***Feasibility threshold.*** In practice, CVX often returns a small positive objective ($r \approx 10^{-7}$) for values of h that are just feasible. Hence the bisection step is accepted if $r < \epsilon$ where $\epsilon \ll 1$. The results are generally insensitive (up to the first few digits) to the choice of ϵ over a large range of values; we have used $\epsilon = 10^{-7}$ for all results in this work. The accuracy that can be achieved is eventually limited by the need to choose a suitable value ϵ.

3.2. ***Conditioning and change of basis.*** Unfortunately, for large values of $h\lambda$, the numerical solution of Problem 2 becomes difficult due to ill-conditioning of the constraint matrix. Observe from (3) that the constrained quantities $R(h\lambda)$ are related to the decision variables a_j through multiplication by a Vandermonde matrix. Vandermonde matrices are known to be ill-conditioned for most choices of abscissas. For very large $h\lambda$, the resulting CVX problem cannot be reliably solved by either of the core solvers.

A first approach to reducing the condition number of the constraint matrix is to rescale the monomial basis. We have found that a more robust approach for many types of spectra can be obtained by choosing a basis that is approximately

orthogonal over the given spectrum $\{\Lambda\}$. Thus we seek a solution of the form

$$R(z) = \sum_{j=0}^{s} a_j Q_j(z), \quad \text{where } Q_j(z) = \sum_{k=0}^{j} b_{jk} z^k. \tag{10}$$

Here $Q_j(z)$ is a degree-j polynomial chosen to give a well-conditioned constraint matrix. The drawback of not using the monomial basis is that the dimension of the problem is $s+1$ (rather than $s+1-p$) and we must now impose the order conditions explicitly:

$$\sum_{j=0}^{s} a_j b_{jk} = \frac{1}{k!} \quad \text{for } k = 0, 1, \ldots, p. \tag{11}$$

Consequently, using a nonmonomial basis increases the number of design variables in the problem and introduces an equality constraint matrix $B \in \mathbb{R}^{p \times s}$ that is relatively small (when $p \ll s$), but usually very poorly conditioned. However, it can dramatically improve the conditioning of the inequality constraints.

The choice of the basis $Q_j(z)$ is a challenging problem in general. In the special case of a negative real spectrum, an obvious choice is the Chebyshev polynomials (of the first kind) T_j, shifted and scaled to the domain $[hx, 0]$ where $x = \min_{\lambda \in \Lambda} \mathrm{Re}(\lambda)$, via an affine map:

$$Q_j(z) = T_j\left(1 + \frac{2z}{hx}\right). \tag{12}$$

The motivation for using this basis is that $|Q_j(h\lambda)| \leq 1$ for all $\lambda \in [hx, 0]$. This basis is also suggested by the fact that $Q_j(z)$ is the optimal stability polynomial in terms of negative real axis inclusion for $p = 1, s = j$. In Section 4, we will see that this choice of basis works well for more general spectra when the largest magnitude eigenvalues lie near the negative real axis.

As an example, we consider a spectrum of 3200 equally spaced values λ in the interval $[-1, 0]$. Figure 2 shows the relative error as well as the condition number of the $3200 \times s$ inequality constraint matrix obtained by using the monomial (3) and Chebyshev (12) bases for $p = 1$ and s ranging from 2 through 50. The optimal objective value is $h = 2s^2$, and the condition number of the inequality constraint matrix is measured for the feasibility problem at this value. The condition number of the monomial basis scales exponentially, while the condition number of the Chebyshev basis constraint matrix has a weak linear dependence on s. Typically, the solver is accurate until the condition number reaches about 10^{16}. This supports the hypothesis that it is the conditioning of the inequality constraint matrix that leads to failure of the solver. The Chebyshev basis keeps the condition number small and yields accurate answers even for very large values of h.

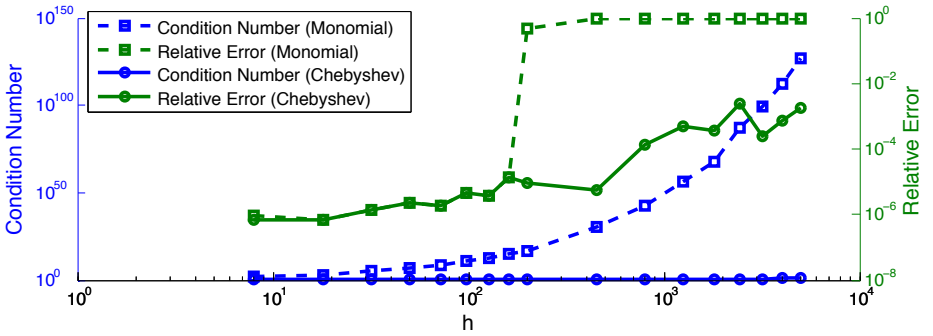


Figure 2. Condition number of principal constraint matrix and relative solution accuracy versus optimal step size. The points along a given curve correspond to different choices of s. The values plotted correspond to $s = 2, 3, \ldots, 9, 10, 15, 20, \ldots, 45, 50$ and a spectrum of 3200 equally spaced values in the interval $[-1, 0]$. The constraint matrix is formed using the optimal value $h = 2s^2$. The Chebyshev basis keeps the condition number small and yields accurate answers even for very large values of h.

4. Examples

We now demonstrate the effectiveness of our algorithm by applying it to determine optimally stable polynomials (i.e., solve Problem 1) for various types of spectra. As stated above, we use Algorithm 1 for its simplicity, speed, and effectiveness. When Λ corresponds to an infinite set, we approximate it by a fine discretization.

4.1. ***Verification.*** In this section, we apply our algorithm to some well-studied cases with known exact or approximate results in order to verify its accuracy and correctness. In addition to the real axis, imaginary axis, and disk cases below, we have successfully recovered the results of [28]. Our algorithm succeeds in finding the globally optimal solution in every case for which it is known, except in some cases of extremely large step sizes for which the underlying solvers (SDPT3 and SeDuMi) eventually fail.

Negative real axis inclusion. Here we consider the largest h such that $[-h, 0] \in S$ by taking $\Lambda = [-1, 0]$. This is the most heavily studied case in the literature, as it applies to the semidiscretization of parabolic PDEs and a large increase of H_{opt} is possible when s is increased (see, e.g., [33; 44; 27; 4; 35]). For first-order accurate methods ($p = 1$), the optimal polynomials are just shifted Chebyshev polynomials, and the optimal timestep is $H_{\text{opt}} = 2s^2$. Many special analytical and numerical techniques have been developed for this case; the most powerful seems to be that of Bogatyrev [4].

We apply our algorithm to a discretization of Λ (using 6400 evenly spaced points) and using the shifted and scaled Chebyshev basis (12). Results for up to $s = 40$ are shown in Table 1 (note that we list H_{opt}/s^2 for easy comparison,

	H_{opt}/s^2				
Stages	$p=1$	$p=2$	$p=3$	$p=4$	$p=10$
1	2.000				
2	2.000	0.500			
3	2.000	0.696	0.279		
4	2.000	0.753	0.377	0.174	
5	2.000	0.778	0.421	0.242	
6	2.000	0.792	0.446	0.277	
7	2.000	0.800	0.460	0.298	
8	2.000	0.805	0.470	0.311	
9	2.000	0.809	0.476	0.321	
10	2.000	0.811	0.481	0.327	0.051
15	2.000	0.817	0.492	0.343	0.089
20	2.000	0.819	0.496	0.349	0.120
25	2.000	0.820	0.498	0.352	0.125
30	2.001	0.821	0.499	0.353	0.129
35	2.000	0.821	0.499	0.354	0.132
40	2.000	0.821	0.500	0.355	0.132

Table 1. Scaled size of real axis interval inclusion for optimized methods.

since H_{opt} is approximately proportional to s^2 in this case). We include results for $p = 10$ to demonstrate the algorithm's ability to handle high-order methods. For $p = 1$ and 2, the values computed here match those available in the literature [43]. Most of the values for $p = 3, 4$ and 10 are new results. Figure 3 shows some examples of stability regions for optimal methods. As observed in the literature, it seems that H_{opt}/s^2 tends to a constant (that depends only on p) as s increases. For large values of s, some results in the table have an error of about 10^{-3} due to inaccuracies in the numerical results provided by the interior point solvers.

Imaginary axis inclusion. Next we consider the largest h such that $[-ih, ih] \in S$ by taking $\Lambda = xi$, $x \in [-1, 1]$. Optimal polynomials for imaginary axis inclusion

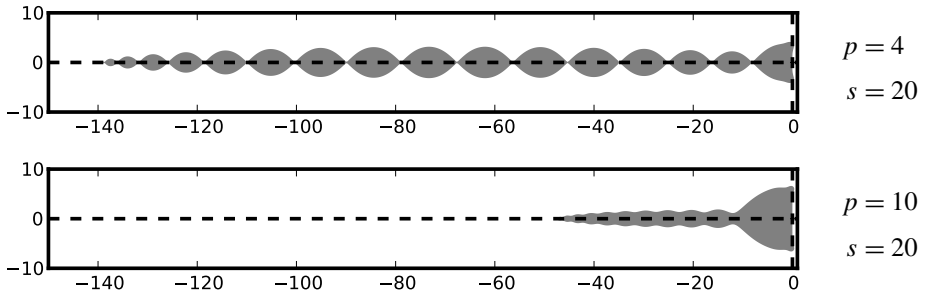


Figure 3. Stability regions of some optimal methods for real axis inclusion.

	H_{opt}/s			
Stages	$p=1$	$p=2$	$p=3$	$p=4$
2	0.500			
3	0.667	0.667	0.577	
4	0.750	0.708	0.708	0.707
5	0.800	0.800	0.783	0.693
6	0.833	0.817	0.815	0.816
7	0.857	0.857	0.849	0.813
8	0.875	0.866	0.866	0.866
9	0.889	0.889	0.884	0.864
10	0.900	0.895	0.895	0.894
15	0.933	0.933	0.932	0.925
20	0.950	0.949	0.949	0.949
25	0.960	0.960	0.959	0.957
30	0.967	0.966	0.966	0.966
35	0.971	0.971	0.971	0.970
40	0.975	0.975	0.975	0.975
45	0.978	0.978	0.978	0.977
50	0.980	0.980	0.980	0.980

Table 2. Scaled size of imaginary axis inclusion for optimized methods.

have also been studied by many authors, and a number of exact results are known or conjectured [43; 46; 20; 21; 22; 44]. We again approximate the problem, taking $N = 3200$ evenly spaced values in the interval $[0, i]$ (note that stability regions are necessarily symmetric about the real axis since $R(z)$ has real coefficients). We use a "rotated" Chebyshev basis defined by

$$Q_j(z) = i^j T_j\left(\frac{iz}{hx}\right),$$

where $x = \max_i(|\operatorname{Im}(\lambda_i)|)$. Like the Chebyshev basis for the negative real axis, this basis dramatically improves the robustness of the algorithm for imaginary spectra. Table 2 shows the optimal effective step sizes. In agreement with [43; 21], we find $H = s - 1$ for $p = 1$ (all s) and for $p = 2$ (s odd). We also find $H = s - 1$ for $p = 1$ and s even, which was conjectured in [46] and confirmed in [44]. We find

$$H_{\text{opt}} = \sqrt{s(s-2)}$$

for $p = 2$ and s even, strongly suggesting that the polynomials given in [20] are optimal for these cases; on the other hand, our results show that those polynomials, while third order accurate, are not optimal for $p = 3$ and s odd. Figure 4 shows some examples of stability regions for optimal methods.

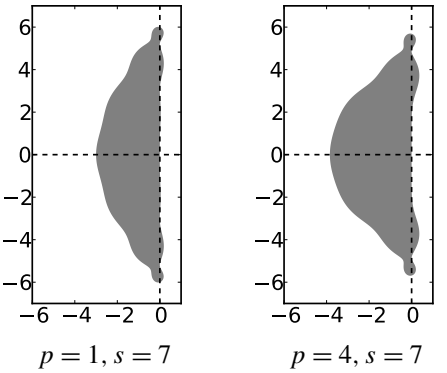


Figure 4. Stability regions of some optimal methods for imaginary axis inclusion.

Disk inclusion. In the literature, attention has been paid to stability regions that include the disk

$$D(h) = \{z : |1 + z/h| \leq 1\}, \tag{13}$$

for the largest possible h. As far as we know, the optimal result for $p = 1$ ($H_{\text{opt}} = s$) was first proved in [15]. The optimal result for $p = 2$ ($H_{\text{opt}} = s - 1$) was first proved in [46]. Both results have been unwittingly rediscovered by later authors. For $p > 2$, no exact results are available.

We use the basis

$$Q_j(z) = \left(1 + \frac{z}{h}\right)^j.$$

Note that $Q_j(z)$ is the optimal polynomial for the case $s = j$, $p = 1$. This basis can also be motivated by recalling that Vandermonde matrices are perfectly conditioned when the points involved are equally spaced on the unit circle. Our basis can be obtained by taking the monomial basis and applying an affine transformation that shifts the unit circle to the disk (13). This basis greatly improves the robustness of

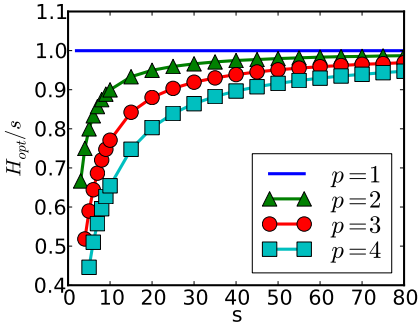


Figure 5. Relative size of largest disk that can be included in the stability region (scaled by the number of stages).

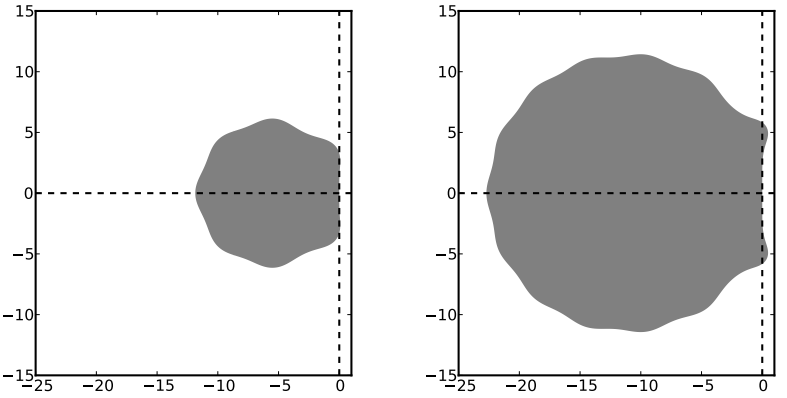


Figure 6. Stability regions of some optimal methods for disk inclusion.

the algorithm for this particular spectrum. We show results for $p \leq 4$ in Figure 5. For $p = 3$ and $s = 5, 6$, our results give a small improvement over those of [16]. Some examples of optimal stability regions are plotted in Figure 6.

4.2. ***Spectrum with a gap.*** We now demonstrate the effectiveness of our method for more general spectra. First we consider the case of a dissipative problem with two time scales, one much faster than the other. This type of problem was the motivation for the development of projective integrators in [8]. Following the ideas outlined there we consider

$$\Lambda = \{z : |z| = 1, \mathbb{R}(z) \leq 0\} \cup \{z : |z - \alpha| = 1\}. \tag{14}$$

We take $\alpha = 20$ and use the shifted and scaled Chebyshev basis (12). Results are shown in Figures 7 and 8. A dramatic increase in efficiency is achieved by adding a few extra stages.

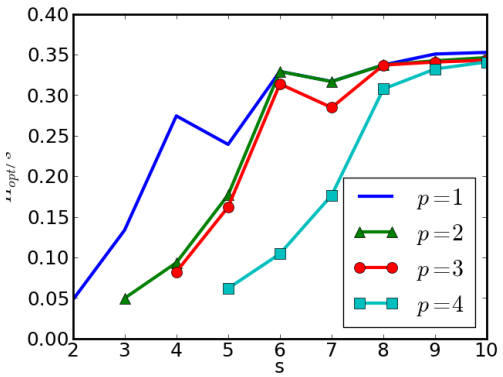


Figure 7. Optimal effective step size for spectrum with a gap (14) with $\alpha = 20$.

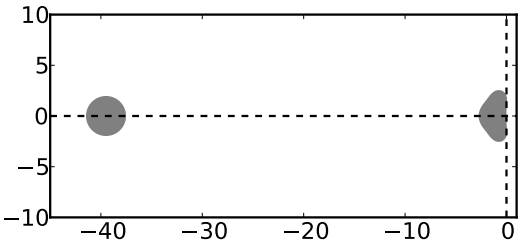


Figure 8. Optimal stability region for $p = 1, s = 6, \alpha = 20$ (stable step size ≈ 1.975).

4.3. ***Legendre pseudospectral discretization.*** Next we consider a system obtained from semidiscretization of the advection equation on the interval $[-1, 1]$ with homogeneous Dirichlet boundary condition:

$$u_t = u_x, \qquad u(t, x = 1) = 0.$$

The semidiscretization is based on pseudospectral collocation at points given by the zeros of the Legendre polynomials; we take $N = 50$ points. The semidiscrete system takes the form (1), where L is the Legendre differentiation matrix, whose eigenvalues are shown in Figure 9(a). We compute an optimally stable polynomial based on the spectrum of the matrix, taking $s = 7$ and $p = 1$. The stability region of the resulting method is plotted in Figure 9(c). Using an appropriate step size, all the scaled eigenvalues of L lie in the stability region. However, this method is unstable in practice for any positive step size; Figure 9(e) shows an example of a computed solution after three steps, where the initial condition is a Gaussian. The resulting instability is nonmodal, meaning that it does not correspond to any of the eigenvectors of L (compare [41, Figure 31.2]).

This discretization is now well-known as an example of nonnormality [41, Chapters 30–32]. Due to the nonnormality, it is necessary to consider pseudospectra in order to design an appropriate integration scheme. The ϵ-pseudospectrum (see [41]) is the set

$$\{z \in \mathbb{C} : \|(zI - L)^{-1}\|_2 > 1/\epsilon\}.$$

The ϵ-pseudospectrum (for $\epsilon = 2$) is shown with the eigenvalues in Figure 9(b); note that the pseudospectrum includes small islands around the isolated eigenvalues, though they are not visible at the scale plotted. The instability observed above occurs because the stability region does not contain an interval on the imaginary axis about the origin, whereas the pseudospectrum includes such an interval.

We now compute an optimally stable integrator based on the 2-pseudospectrum. This pseudospectrum is computed using an approach proposed in [40, Section 20], with sampling on a fine grid. In order to reduce the number of constraints and speed up the solution, we compute the convex hull of the resulting set and apply

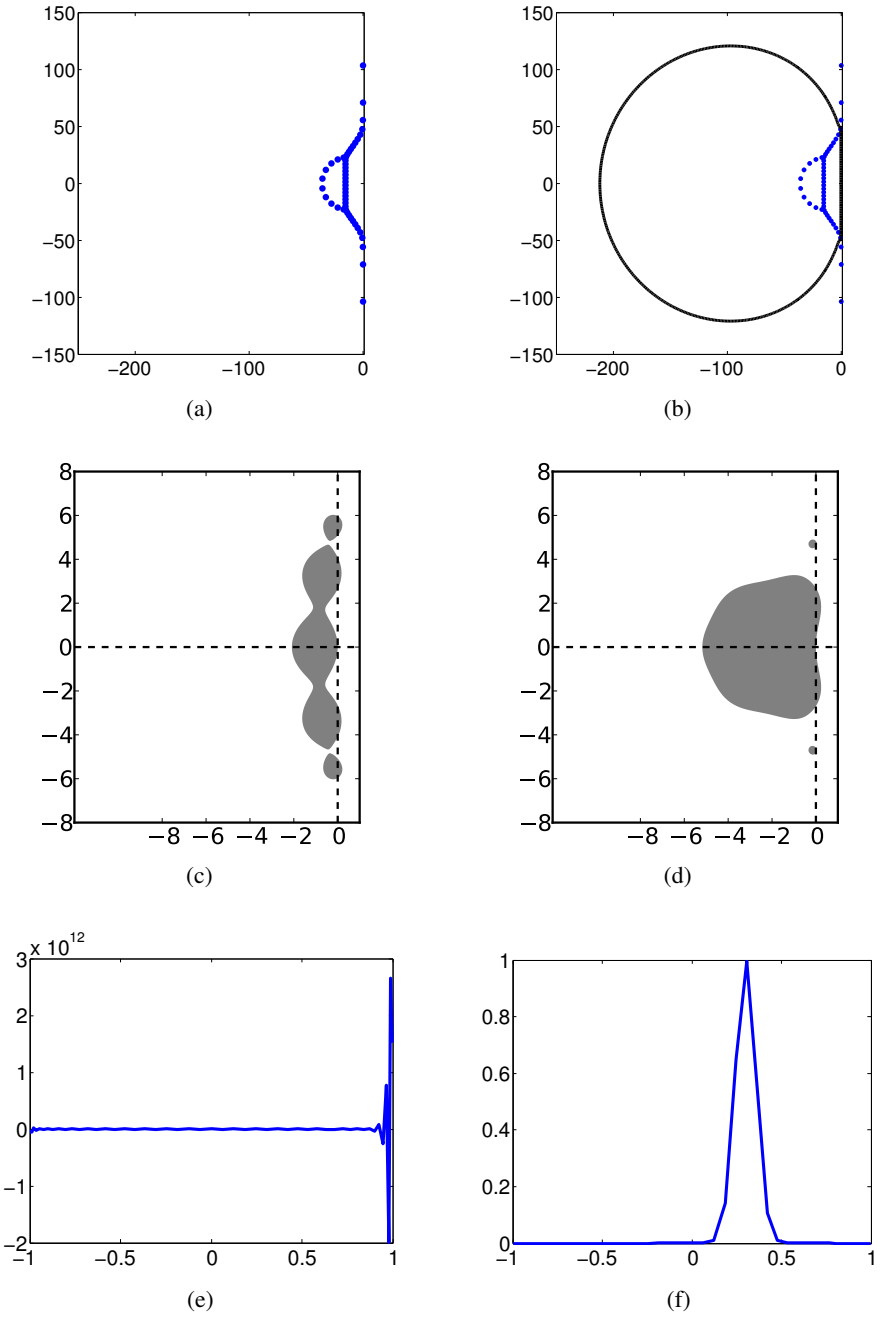


Figure 9. Results for the Legendre differentiation matrix with $N = 50$. Top row: eigenvalues (a) and eigenvalues with pseudospectrum (b). The boundary of the 2-pseudospectrum is plotted. Middle row: Optimized stability region based on eigenvalues (c) and on the pseudospectrum (d). Bottom row: Solution computed with method based on spectrum (e) and with method based on pseudospectrum (f).

our algorithm. The resulting stability region is shown in Figure 9(d). It is remarkably well adapted; notice the two isolated roots that ensure stability of the modes corresponding to the extremal imaginary eigenvalues. We have verified that this method produces a stable solution, in agreement with theory (see Chapter 32 of [41]); Figure 9(f) shows an example of a solution computed with this method. The initial Gaussian pulse advects to the left.

4.4. ***Thin rectangles.*** A major application of explicit Runge–Kutta methods with many stages is the solution of moderately stiff advection-reaction-diffusion problems [14; 45]. For such problems, the stability region must include not only a large interval on the negative real axis, but also some region around it, due to convective terms. If centered differences are used for the advective terms, it is natural to require that a small interval on the imaginary axis be included. Hence, one may be interested in methods that contain a rectangular region

$$\Lambda_\kappa = \{\lambda \in \mathbb{C} : -\beta \le \mathrm{Im}(\lambda) \le \beta,\ -\kappa \le \mathrm{Re}(\lambda) \le 0\}. \tag{15}$$

for given κ, β. Most methods in the literature do not satisfy this requirement (with the notable exception of those in [47]. Most available approaches for devising methods with extended real axis stability (including those of [38]) cannot be applied to such regions. Because of this, most existing methods are applicable only if upwind differencing is applied to convective terms [45; 38].

For this example, rather than parametrizing by the step size h, we assume that a desired step size h and imaginary axis limit β are given based on the convective terms, which generally require small step sizes for accurate resolution. We seek to find (for given s, p) the polynomial (3) that includes Λ_κ for κ as large as possible. This could correspond to selection of an optimal integrator based on the ratio of convective and diffusive scales (roughly speaking, the Reynolds number). Since the desired stability region lies relatively near the negative real axis, we use the shifted and scaled Chebyshev basis (12).

Stability regions of some optimal methods are shown in Figure 10. The outline of the included rectangle is superimposed in black. The stability region for $\beta = 10, s = 20$, shown in Figure 10 is especially interesting as it is very nearly rectangular. A closeup view of the upper boundary is shown in Figure 11. These regions appear to satisfy the hypothesis stated in [38] that their boundary is tangent to the prescribed boundary at $s - p$ points in the upper half plane.

5. Discussion

The approach described here can speed up the integration of IVPs for which

- explicit Runge–Kutta methods are appropriate;

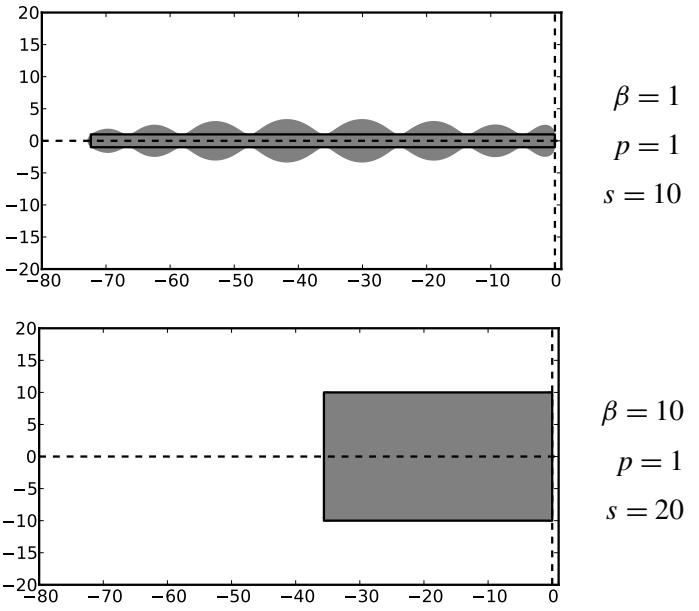


Figure 10. Stability regions of some optimal methods for thin rectangle inclusion.

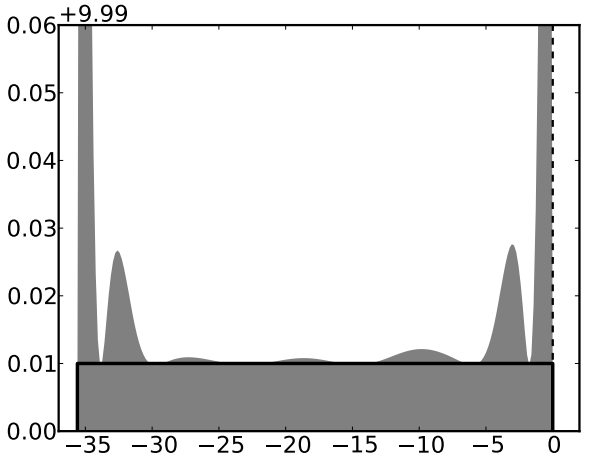


Figure 11. Closeup view of upper boundary of the rectangular stability region plotted in Figure 10.

- the spectrum of the problem is known or can be approximated; and
- stability is the limiting factor in choosing the step size.

Although we have considered only linear initial value problems, we expect our approach to be useful in designing integrators for nonlinear problems via the usual approach of considering the spectrum of the Jacobian. A first successful application of our approach to nonlinear PDEs appears in [30].

The amount of speedup depends strongly on the spectrum of the problem, and can range from a few percent to several times or more. Based on past work and on results presented in Section 4, we expect that the most substantial gains in efficiency will be realized for systems whose spectra have large negative real parts, such as for semidiscretization of PDEs with significant diffusive or moderately stiff reaction components. As demonstrated in Section 4, worthwhile improvements may also be attained for general systems, and especially for systems whose spectrum contains gaps.

The work presented here suggests several extensions and areas for further study. For very high polynomial degree, the convex subproblems required by our algorithm exhibit poor numerical conditioning. We have proposed a first improvement by change of basis, but further improvements in this regard could increase the robustness and accuracy of the algorithm. It seems likely that our algorithm exhibits global convergence in general circumstances beyond those for which we have proven convergence. The question of why bisection seems to always lead to globally optimal solutions merits further investigation. While we have focused primarily on design of the stability properties of a scheme, the same approach can be used to optimize accuracy efficiency, which is a focus of future work. Our algorithm can also be applied in other ways; for instance, it could be used to impose a specific desired amount of dissipation for use in multigrid or as a kind of filtering.

One of the most remarkable aspects of our algorithm is its speed, which opens up the potential for a new kind of adaptive time stepping in which the time integration method itself is designed on-the-fly during the computation. For nonlinear problems, the method could be adapted, for instance, when a significant change in the spectrum of the linearized semidiscretization is detected. Whereas traditional automatic integrators dynamically adjust the step size and scheme order, choosing from a small set of preselected methods, our algorithm could be used as the basis for an implementation that also automatically adjusts details of the stability polynomial at each step. Practical implementation of this idea is dependent on the ability to efficiently approximate this spectrum and might require an implementation of our algorithm in a compiled language.

The problem of determining optimal polynomials subject to convex constraints is very general. Convex optimization techniques have already been exploited to solve similar problems in filter design [7], and will likely find further applications in numerical analysis.

Companion website

The codes used in producing the numerical results in this paper are available at http://www.runmycode.org/CompanionSite/Site158 [18].

Acknowledgments

We thank Lajos Lóczi for providing a simplification of the proof of Lemma 3. We are grateful to R. J. LeVeque and L. N. Trefethen for helpful comments on a draft of this work. We thank the anonymous referees for their comments that improved this paper.

References

[1] A. Abdulle, *On roots and error constants of optimal stability polynomials*, BIT **40** (2000), no. 1, 177–182. MR 2001a:65080 Zbl 0956.65068

[2] ———, *Fourth order Chebyshev methods with recurrence relation*, SIAM J. Sci. Comput. **23** (2002), no. 6, 2041–2054. MR 2003g:65074 Zbl 1009.65048

[3] A. Abdulle and A. A. Medovikov, *Second order Chebyshev methods based on orthogonal polynomials*, Numer. Math. **90** (2001), no. 1, 1–18. MR 2002i:65071 Zbl 0997.65094

[4] A. B. Bogatyrev, *Effective solution of the problem of the best stability polynomial*, Mat. Sb. **196** (2005), no. 7, 27–50, In Russian; translated in Sbornik: Math. **196** (2005), 959–981. MR 2007b:34124a

[5] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, Cambridge, 2004. MR 2005d:90002 Zbl 1058.90049

[6] J. C. Butcher, *Numerical methods for ordinary differential equations*, 2nd ed., Wiley, Chichester, 2008. MR 2009b:65002 Zbl 1167.65041

[7] T. Davidson, *Enriching the art of FIR filter design via convex optimization*, IEEE Signal Proc. Mag. **27** (2010), 89–101.

[8] C. W. Gear and I. G. Kevrekidis, *Projective methods for stiff differential equations: problems with gaps in their eigenvalue spectrum*, SIAM J. Sci. Comput. **24** (2003), no. 4, 1091–1106. MR 2004c:65065 Zbl 1034.65056

[9] S. Gottlieb, D. Ketcheson, and C.-W. Shu, *Strong stability preserving Runge–Kutta and multistep time discretizations*, World Scientific, Hackensack, NJ, 2011. MR 2012f:65107 Zbl 1241.65064

[10] M. C. Grant and S. P. Boyd, *Graph implementations for nonsmooth convex programs*, Recent advances in learning and control (V. Blondel, S. Boyd, , and H. Kimura, eds.), Lecture Notes in Control and Inform. Sci., no. 371, Springer, London, 2008, pp. 95–110. MR 2409077 Zbl 1205.90223

[11] ———, *CVX: MATLAB software for disciplined convex programming*, Tech. report, 2011.

[12] E. Hairer and G. Wanner, *Solving ordinary differential equations, II: Stiff and differential-algebraic problems*, 2nd ed., Springer Series in Computational Mathematics, no. 14, Springer, Berlin, 1996. MR 97m:65007 Zbl 0859.65067

[13] R. Hettich and K. O. Kortanek, *Semi-infinite programming: theory, methods, and applications*, SIAM Rev. **35** (1993), no. 3, 380–429. MR 94g:90152 Zbl 0784.90090

[14] W. Hundsdorfer and J. Verwer, *Numerical solution of time-dependent advection-diffusion-reaction equations*, Springer Series in Computational Mathematics, no. 33, Springer, Berlin, 2003. MR 2004g:65001 Zbl 1030.65100

[15] R. Jeltsch and O. Nevanlinna, *Largest disk of stability of explicit Runge–Kutta methods*, BIT **18** (1978), no. 4, 500–502. MR 80b:65099 Zbl 0399.65051

[16] R. Jeltsch and M. Torrilhon, *Flexible stability domains for explicit Runge–Kutta methods*, Some topics in industrial and applied mathematics (R. Jeltsch, T.-T. Li, and I. H. Sloan, eds.), Ser. Contemp. Appl. Math. CAM, no. 8, Higher Ed. Press, Beijing, 2007, pp. 152–180. MR 2008m:65180 Zbl 1171.65415

[17] C. A. Kennedy, M. H. Carpenter, and R. M. Lewis, *Low-storage, explicit Runge–Kutta schemes for the compressible Navier–Stokes equations*, Appl. Numer. Math. **35** (2000), no. 3, 177–219. MR 2001k:65111 Zbl 0986.76060

[18] D. I. Ketcheson and A. J. Ahmadia, *Optimal stability polynomials for numerical integration of initial value problems*, Tech. report, 2012.

[19] D. I. Ketcheson and M. Parsani, *RK-opt: Software for the design of optimal runge–kutta methods*, Tech. report, 2012.

[20] I. P. E. Kinnmark and W. G. Gray, *One step integration methods of third-fourth order accuracy with large hyperbolic stability limits*, Math. Comput. Simulation **26** (1984), no. 3, 181–188. MR 85k:65069 Zbl 0539.65051

[21] ———, *One step integration methods with maximum stability regions*, Math. Comput. Simulation **26** (1984), no. 2, 87–92. MR 85f:65079 Zbl 0539.65050

[22] ———, *Fourth-order accurate one-step integration methods with large imaginary stability limits*, Numer. Methods Partial Differential Equations **2** (1986), no. 1, 63–70. MR 89b:65175 Zbl 0623.65077

[23] J. D. Lawson, *An order five Runge–Kutta process with extended region of stability*, SIAM J. Numer. Anal. **3** (1966), 593–597. MR 35 #7589 Zbl 0154.40602

[24] R. J. LeVeque, *Finite difference methods for ordinary and partial differential equations: Steady-state and time-dependent problems*, Soc. Industrial and Applied Math., Philadelphia, PA, 2007. MR 2009a:65173

[25] J. Martín-Vaquero and B. Janssen, *Second-order stabilized explicit Runge–Kutta methods for stiff problems*, Comput. Phys. Comm. **180** (2009), no. 10, 1802–1810. MR 2678453 Zbl 1197.65006

[26] J. L. Mead and R. A. Renaut, *Optimal Runge–Kutta methods for first order pseudospectral operators*, J. Comput. Phys. **152** (1999), no. 1, 404–419. MR 2000a:65083 Zbl 0935.65100

[27] A. A. Medovikov, *High order explicit methods for parabolic equations*, BIT **38** (1998), no. 2, 372–390. MR 99i:65096 Zbl 0909.65060

[28] J. Niegemann, R. Diehl, and K. Busch, *Efficient low-storage Runge–Kutta schemes with optimized stability regions*, J. Comput. Phys. **231** (2012), no. 2, 364–372. MR 2012m:65202 Zbl 1243.65113

[29] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity*, Prentice-Hall, Englewood Cliffs, NJ, 1982, Reprinted Dover, Mineola, NY, 1998. MR 84k:90036 Zbl 0503.90060

[30] M. Parsani, D. I. Ketcheson, and W. Deconinck, *Optimized explicit Runge–Kutta schemes for the spectral difference method applied to wave propagation problems*, preprint, 2012, to appear in SIAM J. Sci. Comput. arXiv 1207.5830

[31] J. Pike and P. L. Roe, *Accelerated convergence of Jameson's finite-volume Euler scheme using van der Houwen integrators*, Comput. & Fluids **13** (1985), no. 2, 223–236. MR 87d:65095 Zbl 0571.76003

[32] R. A. Renaut, *Two-step Runge–Kutta methods and hyperbolic partial differential equations*, Math. Comp. **55** (1990), no. 192, 563–579. MR 91d:65128 Zbl 0724.65076

[33] W. Riha, *Optimal stability polynomials*, Computing (Arch. Elektron. Rechnen) **9** (1972), 37–43. MR 47 #4450 Zbl 0234.65076

[34] J. M. Sanz-Serna and M. N. Spijker, *Regions of stability, equivalence theorems and the Courant–Friedrichs–Lewy condition*, Numer. Math. **49** (1986), no. 2-3, 319–329. MR 87i:65140 Zbl 0574.65106

[35] L. M. Skvortsov, *Explicit stabilized Runge–Kutta methods*, Zh. Vychisl. Mat. Mat. Fiz. **51** (2011), no. 7, 1236–1250, In Russian: translated in Comput. Math. and Math. Phys. **51** (2011), 1153–1166. MR 2906150 Zbl 1249.65156

[36] B. P. Sommeijer and J. G. Verwer, *On stabilized integration for time-dependent PDEs*, J. Comput. Phys. **224** (2007), no. 1, 3–16. MR 2008e:65263 Zbl 1119.65382

[37] J. F. Sturm, *Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw. **11/12** (1999), no. 1-4, 625–653. MR 1778433

[38] M. Torrilhon and R. Jeltsch, *Essentially optimal explicit Runge–Kutta methods with application to hyperbolic-parabolic equations*, Numer. Math. **106** (2007), no. 2, 303–334. MR 2008d:65074 Zbl 1113.65074

[39] T. Toulorge and W. Desmet, *Optimal Runge–Kutta schemes for discontinuous Galerkin space discretizations applied to wave propagation problems*, J. Comput. Phys. **231** (2012), no. 4, 2067–2091. MR 2012m:65353 Zbl 1242.65190

[40] L. N. Trefethen, *Computation of pseudospectra*, Acta numerica, 1999, Acta Numer., no. 8, Cambridge Univ. Press, Cambridge, 1999, pp. 247–295. MR 2002b:65062 Zbl 0945.65039

[41] L. N. Trefethen and M. Embree, *Spectra and pseudospectra: The behavior of nonnormal matrices and operators*, Princeton University Press, Princeton, NJ, 2005. MR 2006d:15001

[42] R. H. Tütüncü, K. C. Toh, and M. J. Todd, *Solving semidefinite-quadratic-linear programs using SDPT3*, Math. Program. **95** (2003), no. 2, Ser. B, 189–217. MR 2004c:90036 Zbl 1030.90082

[43] P. J. van der Houwen, *Explicit Runge–Kutta formulas with increased stability boundaries*, Numer. Math. **20** (1972), 149–164. MR 47 #6094 Zbl 0233.65039

[44] ———, *The development of Runge–Kutta methods for partial differential equations*, Appl. Numer. Math. **20** (1996), no. 3, 261–272. MR 97f:65053 Zbl 0857.65094

[45] J. G. Verwer, B. P. Sommeijer, and W. Hundsdorfer, *RKC time-stepping for advection-diffusion-reaction problems*, J. Comput. Phys. **201** (2004), no. 1, 61–79. MR 2005h:65151 Zbl 1059.65085

[46] R. Vichnevetsky, *New stability theorems concerning one-step numerical methods for ordinary differential equations*, Math. Comput. Simulation **25** (1983), no. 3, 199–205. MR 85b:65082 Zbl 0573.65052

[47] C. J. Zbinden, *Partitioned Runge–Kutta–Chebyshev methods for diffusion-advection-reaction problems*, SIAM J. Sci. Comput. **33** (2011), no. 4, 1707–1725. MR 2012m:65208 Zbl 1245.65120

Received July 12, 2012. Revised November 23, 2012.

DAVID I. KETCHESON: david.ketcheson@kaust.edu.sa
Division of Mathematical and Computer Sciences and Engineering, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

ARON J. AHMADIA: aron@ahmadia.net
Division of Mathematical and Computer Sciences and Engineering, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at msp.berkeley.edu/camcos.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in CAMCoS are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

Communications in Applied Mathematics and Computational Science

vol. 7 no. 2 2012

1559-3940(2012)7:2;1-3