The doubly stochastic advection-diffusion-decay model: numerical scheme

Michael Tsyrulnikov and Alexander Rakitko

(michael.tsyrulnikov@gmail.com)

April 18, 2019

1 Model

$$\frac{\partial \xi(t,x)}{\partial t} + U(t,x)\frac{\partial \xi(t,x)}{\partial x} + \rho(t,x)\xi(t,x) - \nu(t,x)\frac{\partial^2 \xi(t,x)}{\partial x^2} = \sigma(t,x)\alpha(t,x). \tag{1}$$

2 Discretization

In an implicit scheme, in which the time derivative $\partial\xi/\partial t$ is approximated as $(\xi_k - \xi_{k-1})/\Delta t$, all other occurrences of ξ (that is, $\partial\xi/\partial x$ ξ, and $\partial^2\xi/\partial x^2$) are evaluated at the current time step k. We also choose to evaluate the coefficients at the current time step k. This leads to the finite-difference scheme

$$\frac{\boldsymbol{\xi}^k - \boldsymbol{\xi}^{k-1}}{\Delta t} + \mathsf{diag}\,(\mathbf{U}^k)\mathbf{D}^{upwind}\boldsymbol{\xi}^k + \mathsf{diag}\,(\boldsymbol{\rho}^k)\boldsymbol{\xi}^k - \mathsf{diag}\,(\boldsymbol{\nu}^k)\mathbf{D_2}\boldsymbol{\xi}^k = \mathsf{diag}\,(\boldsymbol{\sigma}^k)\boldsymbol{\alpha}^k, \tag{2}$$

where the superscript k labels the time step, $\mathbf{D}^{upwind}$ is the upwind non-centered (directed) finite difference operator (matrix) (upwind means that the second point in the stencil besides the central point is the point towards the direction from which the wind blows, see subsection 2.1 below), $\mathbf{D_2}$ is the usual three-point approximation of the Laplacian in 1D (a matrix) such that the action of $\mathbf{D_2}$ on the vector $\mathbf{y}$ returns

$$(\mathbf{D_2}\mathbf{y})_j = \frac{y_{j-1} - 2y_j + y_{j+1}}{(\Delta x)^2} \tag{3}$$

(where the subscript j labels the spatial grid points), $\boldsymbol{\alpha}_j^k = \frac{1}{\sqrt{\Delta t \Delta x}}N(0,1)$ is the time and space discretized white noise, and $\mathsf{diag}\,(\mathbf{z})$ is the diagonal matrix with the vector $\mathbf{z}$ on its main diagonal.

2.1 Upwind finite difference operator

If $U_j > 0$, then the jth row of the matrix $\mathbf{D}^{upwind}$ has the following two non-zero entries:

$$\mathbf{D}^{upwind}_{j,j-1} = -\frac{1}{\Delta x} \tag{4}$$

$$\mathbf{D}^{upwind}_{j,j} = +\frac{1}{\Delta x} \tag{5}$$

If $U_j < 0$, then the jth row of the matrix $\mathbf{D}^{upwind}$ has the following two non-zero entries:

$$\mathbf{D}^{upwind}_{j,j} = -\frac{1}{\Delta x}$$

$$\mathbf{D}^{upwind}_{j,j+1} = +\frac{1}{\Delta x}$$

3 Time stepping

We rewrite Eq.(2) as

$$\mathbf{G}_k \cdot \boldsymbol{\xi}^k = \boldsymbol{\xi}^{k-1} + \Delta t\, \mathsf{diag}\left(\boldsymbol{\sigma}^k\right) \cdot \boldsymbol{\alpha}^k, \tag{6}$$

where

$$\boxed{\mathbf{G}_k = \mathbf{I} + \Delta t\, \mathsf{diag}\left(\mathbf{U}^k\right) \cdot \mathbf{D}^{upwind} + \Delta t\, \mathsf{diag}\left(\boldsymbol{\rho}^k\right) - \Delta t\, \mathsf{diag}\left(\boldsymbol{\nu}^k\right) \cdot \mathbf{D_2}.} \tag{7}$$

So,

$$\boxed{\boldsymbol{\xi}^k = \mathbf{F}_k \left[\boldsymbol{\xi}^{k-1} + \Delta t\, \mathsf{diag}\left(\boldsymbol{\sigma}^k\right) \cdot \boldsymbol{\alpha}^k\right],} \tag{8}$$

where

$$\boxed{\mathbf{F}_k = \mathbf{G}_k^{-1}}. \tag{9}$$

4 Technical details of the solution of the system of linear algebraic equations with the cyclic tridiagonal matrix

4.1 The system

Let

$$\mathbf{M} := \frac{1}{\Delta t}\mathbf{G} \tag{10}$$

By $h = \Delta s$ denote the mesh size.

Conventions.

(i) By $i+1$ we always mean $i+1$ if $i < n$ and 1 if $i = n$.

(ii) Likewise, by $i-1$ we mean $i-1$ if $i > 1$ and n if $i = 1$.

(iii) We drop the time index in this section.

(iv) By $\mathbf{a}, \mathbf{b}, \mathbf{c}$ we denote the lower, main, and upper (sub-)diagonals of $\mathbf{M}$, respectively.

(v) We also often drop the row index if this does not cause an ambiguity.

Then, in any row, we have (NB: the indexes are dropped!):

- If $U > 0$:

$$a = -\frac{U}{h} - \frac{\nu}{h^2} \tag{11}$$

$$b = \frac{1}{\Delta t} + \frac{U}{h} + \rho + 2\frac{\nu}{h^2} \tag{12}$$

$$c = -\frac{\nu}{h^2} \tag{13}$$

- If $U \leq 0$:

$$a = -\frac{\nu}{h^2} \tag{14}$$

$$b = \frac{1}{\Delta t} - \frac{U}{h} + \rho + 2\frac{\nu}{h^2} \tag{15}$$

$$c = \frac{U}{h} - \frac{\nu}{h^2} \tag{16}$$

The diagonal dominance of $\mathbf{M}$ is easily seen.

Then, the system to be solved is

$$\boxed{\mathbf{M} \cdot \mathbf{x} = \mathbf{y}} \tag{17}$$

where

$$\mathbf{x} := \boldsymbol{\xi}^k \tag{18}$$

and

$$\mathbf{y} := \frac{\boldsymbol{\xi}^{k-1}}{\Delta t} + \boldsymbol{\sigma}^k \cdot \boldsymbol{\alpha}^k \tag{19}$$

and

$$\boldsymbol{\alpha}^k = \frac{1}{\sqrt{\Delta t \Delta s}} \boldsymbol{\sigma}^k \cdot \mathbf{N}(\mathbf{0}, \mathbf{I}). \tag{20}$$

4.2 Solution

4.2.1 Single forecast

To solve the system of linear algebraic equations with the *cyclic tridiagonal matrix* we reduce it to the system with the ordinary *tridiagonal matrix* (i.e. without the "corners"). The latter (banded) system is then solved using the function `Solve.tridiag` from the R package `limSolve` (Soetaert et al., 2009).

The reduction of the *cyclic tridiagonal matrix* to the *non-cyclic tridiagonal matrix* is performed following Ahlberg et al. (2016, p.15). Specifically,

1. From the matrix $\mathbf{M}$, we remove the first row and the first column. The remainder is the $(n-1) \times (n-1)$ truly banded tridiagonal matrix, let us denote it $\widetilde{\mathbf{M}}$.

2. Then, we remove the first element of the r.h.s. $\mathbf{y}$ and denote the rest by $\widetilde{\mathbf{y}}$, so that

$$\mathbf{y}^\top = (y_1, \widetilde{\mathbf{y}}^\top). \tag{21}$$

3. Correspondingly, we remove the first element of the solution $\mathbf{x}$, denote the rest by $\widetilde{\mathbf{x}}$, and denote x_1 by χ, so that

$$\mathbf{x}^\top = (\chi, \widetilde{\mathbf{x}}^\top). \tag{22}$$

As a result, we obtain

$$\left(\begin{array}{c|ccccc} b_1 & c_1 & 0 & \dots & 0 & a_1 \\ \hline a_2 & b_2 & c_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_n & 0 & 0 & \dots & a_n & b_n \end{array}\right) \cdot \left(\begin{array}{c} \chi \\ \hline x_2 \\ \vdots \\ x_n \end{array}\right) = \left(\begin{array}{c} y_1 \\ \hline y_2 \\ \vdots \\ y_n, \end{array}\right) \tag{23}$$

which is regarded as the system of the two equations. The first equation is

$$\boxed{b_1\chi + c_1 x_2 + a_1 x_n = y_1.} \tag{24}$$

The second equation is

$$\widetilde{\mathbf{M}}\widetilde{\mathbf{x}} + \boldsymbol{\varphi}\chi = \widetilde{\mathbf{y}}, \tag{25}$$

where

$$\boldsymbol{\varphi} := (a_2, 0, \dots, 0, c_n)^\top. \tag{26}$$

Let us rewrite the latter equation as

$$\boxed{\widetilde{\mathbf{M}}\widetilde{\mathbf{x}} = \widetilde{\mathbf{y}} - \boldsymbol{\varphi}\chi} \tag{27}$$

and note that the solution to *linear* Eq.(27) can be found by applying the superposition principle. We write down two systems (with the same non-cyclic tridiagonal matrix $\widetilde{\mathbf{M}}$) in which the right-hand sides come from the two components of the r.h.s. in Eq.(27) as follows.
System 1 is

$$\widetilde{\mathbf{M}}\mathbf{u} = \widetilde{\mathbf{y}}. \tag{28}$$

System 2 is

$$\widetilde{\mathbf{M}}\mathbf{v} = -\boldsymbol{\varphi}. \tag{29}$$

Note that $\mathbf{u}$ and $\mathbf{v}$ can be explicitly found (e.g. using function `Solve.tridiag`). Then, it is easily verified that the solution to the system Eq.(27) is

$$\boxed{\widetilde{\mathbf{x}} = \mathbf{u} + \mathbf{v} \cdot \chi} \tag{30}$$

Thus, due to Eq.(30), the vector $\widetilde{\mathbf{x}}$ is known up to the one single unknown, χ. To find χ, we use Eq.(24). Specifically, from Eq.(30), we express

$$x_2 = u_1 + \chi \cdot v_1 \tag{31}$$

and

$$x_n = u_{n-1} + \chi \cdot v_{n-1} \tag{32}$$

(where, we recall, all elements of the vectors $\mathbf{u}$ and $\mathbf{v}$ are known at this point), so that Eq.(24) becomes

$$b_1\chi + c_1(u_1 + \chi \cdot v_1) + a_1(u_{n-1} + \chi \cdot v_{n-1}) = y_1, \tag{33}$$

whence

$$(b_1 + c_1v_1 + a_1v_{n-1}) \cdot \chi + c_1u_1 + a_1u_{n-1} = y_1 \tag{34}$$

and so

$$\boxed{\chi = \frac{y_1 - c_1u_1 - a_1u_{n-1}}{b_1 + c_1v_1 + a_1v_{n-1}}} \tag{35}$$

Having found χ, we substitute it into Eq.(30), getting $\widetilde{\mathbf{x}}$, and form the final solution $\mathbf{x}$ by concatenating χ and $\widetilde{\mathbf{x}}$ using Eq.(22).

4.2.2 Ensemble of forecasts

If we wish to generate an ensemble of N members that share the secondary (coefficient) fields, then, in the above notation, $\mathbf{x}$ and $\mathbf{y}$ become $n \times N$ matrices, $\mathbf{X}$ and $\mathbf{Y}$, respectively.

Correspondingly, $\widetilde{\mathbf{x}}$, $\widetilde{\mathbf{y}}$, $\mathbf{u}$ become $(n-1) \times N$ matrices, $\widetilde{\mathbf{X}}$, $\widetilde{\mathbf{Y}}$, $\mathbf{U}$ respectively. The vector $\mathbf{v}$ remains a vector. The vector $\boldsymbol{\varphi}$ also remains unchanged.

The scalar χ becomes an N-vector, $\boldsymbol{\chi}$.

Computationally:

(1) System Eq.(28) is straightforwardly solved using `Solve.tridiag` with N right-hand sides,

$$\widetilde{\mathbf{M}}\mathbf{U} = \widetilde{\mathbf{Y}}. \tag{36}$$

(2) Eq.(29) is still solved only once.

(3) Eq.(27) becomes

$$\boxed{\widetilde{\mathbf{M}}\widetilde{\mathbf{X}} = \widetilde{\mathbf{Y}} - \boldsymbol{\varphi} \cdot \boldsymbol{\chi}^\top} \tag{37}$$

(4) Eq.(30) becomes

$$\boxed{\widetilde{\mathbf{X}} = \mathbf{U} + \mathbf{v} \cdot \boldsymbol{\chi}^\top} \tag{38}$$

(5) Eq.(24) becomes an element-wise equation, in which χ, x_2, x_n, y_1 are N-vectors. Correspondingly, Eq.(35) is now understood as a element-wise expression for the N-vectors $\mathbf{y}_1, \mathbf{u}_1, \mathbf{u}_{n-1}$:

$$\boxed{\boldsymbol{\chi} = \frac{\mathbf{y}_1 - c_1\mathbf{u}_1 - a_1\mathbf{u}_{n-1}}{b_1 + c_1v_1 + a_1v_{n-1}}} \tag{39}$$

5 Stability

Statement. *If the coefficient fields ρ and ν are* positive, *then the numerical scheme Eqs.(6)–(9) is* **unconditionally stable** *(i.e. stable for any time step Δt).*

To prove this statement, we note that if $\mathbf{F}_k = \mathbf{F} = \mathsf{const}$, then the stability is equivalent to the *boundedness of norms of powers of the transition operator* $\mathbf{F}_k$ (Godunov and Ryabenki, 1964). If $\mathbf{F}_k \neq \mathsf{const}$, then it is sufficient to prove that $\|\mathbf{F}_1 \cdot \mathbf{F}_2 \cdot \dots \cdot \mathbf{F}_N\| < C$, where N is the total number of time steps and $C > 0$ is a constant. The latter condition is, in turn, satisfied if $\forall k$, $\mathbf{F}_k$ is a *contraction*, that is,

$$\boxed{\|\mathbf{F}_k\| < 1}. \tag{40}$$

Now, we show that this latter condition is indeed met in our case.

To this end, let us introduce the vector norm for the spatial vector $\boldsymbol{\xi}$. Let us select the *maximal* (or l_∞) norm:

$$\|\boldsymbol{\xi}\|_\infty := \max_i |\xi_i|. \tag{41}$$

Note that the maximal vector norm induces the matrix norm

$$\|\mathbf{F}\|_\infty = \max_i \sum_j |F_{ij}|. \tag{42}$$

Then, we show that $\forall \boldsymbol{\xi}$,

$$\|\mathbf{G}_k \boldsymbol{\xi}\|_\infty > \|\boldsymbol{\xi}\|_\infty. \tag{43}$$

This condition would entail that the condition Eq.(40) because if we denote $\boldsymbol{\eta} := \mathbf{G}_k \boldsymbol{\xi}$ and note that $\mathbf{F}_k = \mathbf{G}_k^{-1}$. Indeed. Eq.(43) implies that $\forall \boldsymbol{\eta}$,

$$\|\mathbf{F}_k \boldsymbol{\eta}\|_\infty < \|\boldsymbol{\eta}\|_\infty, \tag{44}$$

which, by definition of the operator norm, would mean that

$$\|\mathbf{F}_k\|_\infty \leq 1. \tag{45}$$

So, to prove that the scheme is unconditionally stable, it remains to prove that the condition Eq.(43) is satisfied with our numerical scheme Eqs.(6)–(9).

Take a vector $\boldsymbol{\xi}$. Find its maximal (in modulus) element, let it be ξ_I. We have $\forall i$

$$\sum_j G_{ij} \xi_j = G_{i,i-1} \xi_{i-1} + G_{i,i} \xi_i + G_{i,i+1} \xi_{i+1} \tag{46}$$

(the time index is dropped till the end of the proof).

In Eq.(46), we have in case $U_i > 0$:

$$G_{i,i-1} = -\frac{\Delta t}{\Delta x} U_i - \frac{\Delta t}{(\Delta x)^2} \nu_i \tag{47}$$

$$G_{i,i} = 1 + \frac{\Delta t}{\Delta x} U_i + \Delta t \rho + 2 \frac{\Delta t}{(\Delta x)^2} \nu_i \tag{48}$$

$$G_{i,i+1} = -\frac{\Delta t}{(\Delta x)^2}\nu_i \tag{49}$$

For $U_i \le 0$, we have similarly

$$G_{i,i-1} = -\frac{\Delta t}{(\Delta x)^2}\nu_i \tag{50}$$

$$G_{i,i} = 1 - \frac{\Delta t}{\Delta x}U_i + \Delta t\rho + 2\frac{\Delta t}{(\Delta x)^2}\nu_i \tag{51}$$

$$G_{i,i+1} = \frac{\Delta t}{\Delta x}U_i - \frac{\Delta t}{(\Delta x)^2}\nu_i \tag{52}$$

We have

$$\|\mathbf{G}\boldsymbol{\xi}\|_\infty = \max_i |\sum_j G_{ij}\xi_j| \ge |\sum_j G_{Ij}\xi_j|, \tag{53}$$

where I (we recall) corresponds to the maximal $|\xi_i|$. Here

$$|\sum_j G_{Ij}\xi_j| = |G_{i,i-1}\xi_{I-1} + G_{i,i}\xi_I + G_{i,i+1}\xi_{I+1}| \tag{54}$$

For concreteness and without loss of generality, let $U_I > 0$ and $\xi_I > 0$. Then, from Eqs.(47)–(49), $G_{I,I} > 0$, $G_{I,I-1} < 0$, $G_{I,I+1} < 0$, and $|\xi_{I+1}| \le |\xi_I|$ and $|\xi_{I-1}| \le |\xi_I|$. This implies that the minimal value of $|\sum_j G_{Ij}\xi_j|$ is attained if $\xi_{I-1} > 0$ and $\xi_{I+1} > 0$. Therefore, finally,

$$\|\mathbf{G}\boldsymbol{\xi}\|_\infty \ge |\sum_j G_{Ij}\xi_j| \ge (G_{i,i-1} + G_{i,i} + G_{i,i+1}) \cdot |\xi_I| = (1 + \Delta t\rho) \cdot \|\boldsymbol{\xi}\|_\infty \ge \|\boldsymbol{\xi}\|_\infty, \tag{55}$$

QED.

Actually, we have proved a little bit more than $\|\mathbf{F}_k\| < 1$. Equation (55) implies that

$$\boxed{\|\mathbf{F}_k\| \le \frac{1}{1 + \Delta t\rho} < 1}. \tag{56}$$

As a result, we have proven that **the scheme is unconditionally stable if $\rho > 0$ and $\nu > 0$.**

If the condition of positivity of ρ or ν is violated, then the scheme can become unstable, and this is intentional (the intermittent instability is an important feature of the DSADM).

6 Model error

By model error in Eq.(1) we mean its r.h.s. $\sigma(t,x)\alpha(t,x)$. Rewrite Eq.(8) as

$$\boldsymbol{\xi}^k = \mathbf{F}\boldsymbol{\xi}^{k-1} + \Delta t\, \mathbf{F}\, \mathsf{diag}\,(\boldsymbol{\sigma}^k) \cdot \boldsymbol{\alpha}^k, \tag{57}$$

so that the time discrete model error is

$$\boxed{\boldsymbol{\varepsilon}_k = \Delta t \cdot \mathbf{F} \cdot \left(\mathsf{diag}\,(\boldsymbol{\sigma}^k) \cdot \boldsymbol{\alpha}^k\right)} \tag{58}$$

Note the application of $\mathbf{F}$.

Now, find the model error covariance matrix. Since

$$\mathsf{Var}\,\boldsymbol{\alpha}_i^k = \frac{1}{\Delta t \cdot \Delta x}$$

we obtain

$$\boxed{\mathbf{Q}_k = \frac{\Delta t}{\Delta x} \cdot \mathbf{F} \cdot (\mathsf{diag}\,(\boldsymbol{\sigma}^k))^2 \cdot \mathbf{F}^\top} \tag{59}$$

From this equation, it follows that when computing $\mathbf{B}^k = \mathbf{F}\mathbf{A}^{k-1}\mathbf{F}^\top + \mathbf{Q}^k$, it suffices to compute

$$\mathbf{A}_+^{k-1} = \mathbf{A}^{k-1} + \frac{\Delta t}{\Delta x} \cdot (\mathsf{diag}\,(\boldsymbol{\sigma}^k))^2 \tag{60}$$

and then compute

$$\mathbf{B}^k = \mathbf{F}\mathbf{A}_+^{k-1}\mathbf{F}^\top. \tag{61}$$

In other words, we should add $\mathbf{Q}^k$ to $\mathbf{A}^{k-1}$ — rather than to $\mathbf{F}\mathbf{A}^{k-1}\mathbf{F}^\top$ (i.e. *before* the forecast rather than *after* the forecast).

References

J. H. Ahlberg, E. N. Nilson, and J. L. Walsh. *The Theory of Splines and Their Applications*, volume 38. Elsevier, 2016.

S. K. Godunov and V. S. Ryabenki. *Difference schemes*. Amsterdam: North Holland, 1964.

K. Soetaert, K. Van den Meersche, and D. van Oevelen. limsolve: Solving linear inverse models. *R package version*, 1(1), 2009. URL `https://CRAN.R-project.org/package=limSolve`.