Model reduction: past and present

Jiaxi Zhao

4th Dec, 2023

Words at the beginning

We will focus on **scientific** time series modeling. Model reduction is related to lots of other terminologies such as modal analysis, reduced-order modeling, etc.

The key feature of time series modeling:

* Stability issue

* Extrapolation or interpolation?

$$\begin{aligned} &\mathbf{X}_1 \to \mathbf{X}_2 \to \cdots \mathbf{X}_n, \\ &t_1 \to \mathbf{X}_{t_1}, t_2 \to \mathbf{X}_{t_2}, \cdots \end{aligned} \tag{1}$$

Reduced-order modeling: Past

- Balanced Proper Orthogonal Decomposition (POD)
- Galerkin projection
- Discrete empirical interpolation method (DEIM)
- Koopman operator inspired methods
- Tensor-based methods[1]

[1] Benner, Peter, et al., eds. Model reduction and approximation: theory and algorithms. Society for Industrial and Applied Mathematics, 2017.

Modal analysis: POD

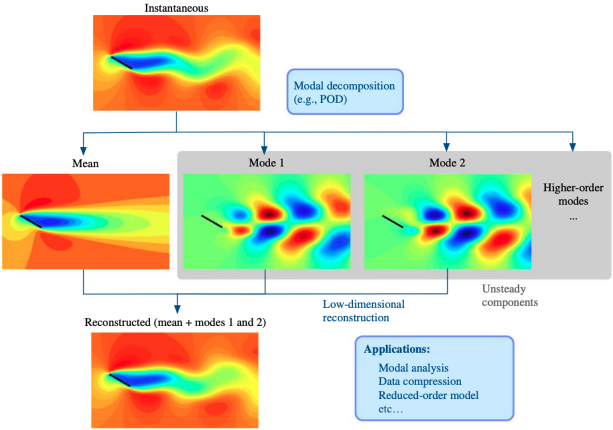

Figure: Modal decomposition of two-dimensional incompressible flow over a flat-plate wing $Re = 100, \alpha = 30$. This example shows complex nonlinear separated flow being well represented by only two POD modes and the mean flowfield. Visualized are the streamwise velocity profiles.[2]

Balanced Transformation

Let us consider the following control system

$$\frac{d}{dt}\mathbf{x}(t) = A\mathbf{x}(t) + B\mathbf{u}(t), \quad \mathbf{y}(t) = C\mathbf{x}(t). \tag{2}$$

The key observation is that any invertible transformation $\widetilde{\mathbf{x}} = V\mathbf{x}$ will result in an equivalent system with different POD basis. For this system, the controllability and observability Grammians are defined as

$$W_c = \int_0^\infty e^{At} BB^T e^{A^T t} dt, \quad W_o = \int_0^\infty e^{A^T t} C^T C e^{At} dt. \tag{3}$$

Balanced transformation V is chosen so that the W_c, W_o are diagonal and equal. [3]

[3] Willcox, Karen, and Jaime Peraire. "Balanced model reduction via the proper orthogonal decomposition." AIAA journal 40.11 (2002): 2323-2330.

Balanced POD

Under the transformation V, two Grammians will transform according to

$$\widetilde{W_c} = V^{-1} W_c V^{-T}, \quad \widetilde{W_o} = V^T W_o V. \tag{4}$$

Then their product transforms as

$$\widetilde{W_c W_o} = V^{-1} W_c W_o V. \tag{5}$$

Projection-based ROM

We consider two types of problem as follows:

$$\begin{aligned} \frac{d}{dt}\mathbf{x}(t) &= A\mathbf{x}(t) + N(\mathbf{x}(t)), \\ 0 &= A_\mu \mathbf{x}(\mu) + N_\mu(\mathbf{x}(\mu)), \quad \mathbf{x} \in \mathbb{R}^{n\times n}. \end{aligned} \tag{6}$$

In both systems, $N(\cdot)$ represents the nonlinearity. Given any reduced basis functions of order k, orthogonal projection operator onto this basis is denoted as V_k with reduced system

$$\begin{aligned} \frac{d}{dt}\widetilde{\mathbf{x}}(t) &= V_k^T A V_k \widetilde{\mathbf{x}}(t) + V_k^T N(V_k \widetilde{\mathbf{x}}(t)), \\ 0 &= V_k^T A_\mu V_k \widetilde{\mathbf{x}}(\mu) + V_k^T N_\mu(V_k \widetilde{\mathbf{x}}(\mu)), \quad \widetilde{\mathbf{x}} \in \mathbb{R}^{k\times n}. \end{aligned} \tag{7}$$

DEIM

The nonlinear term still remains huge amount of computation:

$$V_k^T N(V_k \widetilde{\mathbf{x}}(t)), \quad \widetilde{J}_N(\mathbf{x}(\mu)) = V_k^T J_F(V_k \widetilde{\mathbf{x}}(\mu)) V_k. \tag{8}$$

The idea is to project this nonlinear term further onto a low-dimensional subspace spanned by $\{\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_m\}$ which is obtained by applying POD to the nonlinear snapshots obtained from the original full-order system.

$$N(V_k \widetilde{\mathbf{x}}(t)) = \mathbf{U} c(t). \tag{9}$$

Interpolation method

4

[4] Amsallem, David, and Charbel Farhat. "Interpolation method for adapting reduced-order models and application to aeroelasticity." AIAA journal 46.7 (2008): 1803-1813.

Difficulties of model reduction

* Nonlinearity, e.g. convection

* Transient modeling and unsteady, especially for long time prediction and turbulence

Draw-back of linear-subspace ROM

In particular, linear-subspace ROMs can be expected to produce low-dimensional models with high accuracy[5]only if the problem admits a fast decaying Kolmogorov n-width (e.g., diffusion-dominated problems).

$$d_n(\mathcal{M}) := \inf_{\mathcal{S}_n} \sup_{f} \inf_{g \in \mathcal{S}_n} \|f - g\|. \tag{10}$$

Unfortunately, many problems of interest exhibit a slowly decaying Kolmogorov n-width (e.g., advection-dominated problems).

[5]Binev, Peter, et al. "Convergence rates for greedy algorithms in reduced basis methods." SIAM journal on mathematical analysis 43.3 (2011): 1457-1472.

Koopman operator

Methods related to the Koopman operator are related to the dynamics of the operator, which is also approximated via a linear dynamics

* Extended Dynamical Model Decomposition (EDMD)
* EDMD-DL
* parametric Koopman

ROM: Present

* Nonlinear ROM
* Non-intrusive ROM via operator inference
* Temporal coarsening

Nonlinear trial manifold: learn the reduced basis

Nonlinear trial manifold[6]

$$\widetilde{\mathbf{x}}(t;\mu) = \mathbf{x}_{ref}(\mu) + g(\widehat{\mathbf{x}}(t;\mu)), \tag{11}$$

where $\mathbf{x}_{ref}(\mu)$ denotes the parametrized reference state specified according to the initial condition and $g : \mathbb{R}^p \to \mathbb{R}^n$ denotes the nonlinear parameterization function refered to as *decoder*. The reduced dynamics can be obtained via chain rule:

$$\frac{d}{dt}\widetilde{\mathbf{x}}(t;\mu) = J_g(\widehat{\mathbf{x}}(t;\mu))\frac{d}{dt}\widehat{\mathbf{x}}(t;\mu). \tag{12}$$

[6] Lee, Kookjin, and Kevin T. Carlberg. "Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders." Journal of Computational Physics 404 (2020): 108973.

Time-continuous residual minimization

The model can be written using the residue function

$$\mathbf{r}(\mathbf{v}, \mathbf{x}, t, \mu) = \mathbf{v} - f(\mathbf{x}, t, \mu). \tag{13}$$

Based on this, we can define the equation for the reduced model as

$$\frac{d}{dt}\widehat{\mathbf{x}}(t;\mu) = \arg\min_{\mathbf{v}\in\mathbb{R}^p} \left\|\mathbf{r}(J_g(\widehat{\mathbf{x}}(t;\mu))\mathbf{v}, \mathbf{x}_{ref}(\mu) + g(\widehat{\mathbf{x}}(t;\mu)), t, \mu)\right\| \tag{14}$$

Based on this, the truncation error analysis of the ROM can also be performed using approximation theory of the function spaces.

Operator inference: Learn the reduced operator

Temporal coarsening

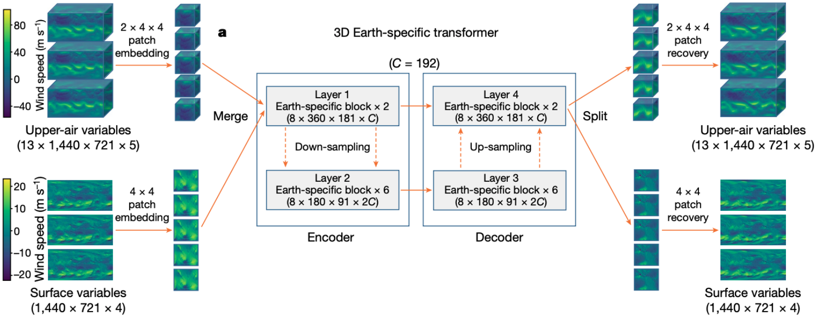

Figure: 3DEST architecture. Based on the standard encoder–decoder design of vision transformers, we adjusted the shifted-window mechanism and applied an Earth-specific positional bias.[7]

[7] Bi, Kaifeng, et al. "Accurate medium-range global weather forecasting with 3D neural networks." Nature 619.7970 (2023): 533-538.

How to do long time prediction?

One of the bottleneck for ROM is the long time prediction accuracy: e.g. for weather forecasting, most data-driven models outperform numerical weather prediction over the 0-7 days regime but quickly

Several methods to perform time series prediction:

* Hierarchical temporal aggregation

* Manifold regularization

* Nonlinear stability issue, especially compared with classical numerical stability

Operator inference ROM

Mesh-based $\Longrightarrow$ Mesh-free

Another kind of nonlinear ROM is based on operator inference. A heuristic: Classical mesh-based solver amounts to solve the high dimesional mapping between the discretization on the huge mesh, e.g. $\mathbb{R}^{N\times N\times N} \to \mathbb{R}^{N\times N\times N}$, how about considering directly $\mathbb{R}^3 \to \mathbb{R}$, which is usually a nonlinear map[8].

Can be viewed as learning the reduced basis and operator simultaneously

[8] Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." Communications of the ACM 65.1 (2021): 99-106.

Operator inference ROM

More over, the parameter can also be fitted into this framework by encoding it as a latent vector[9]

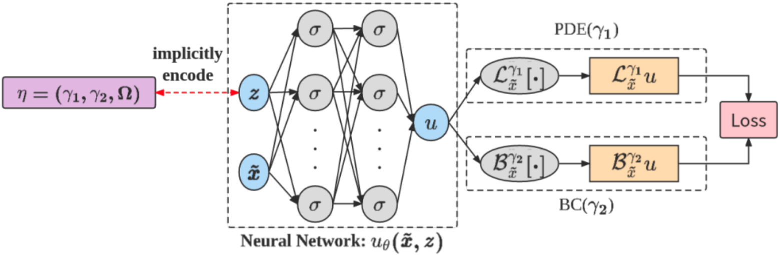

Figure: Architecture of Meta-Auto-Decoder..[10]

[10] Park, Jeong Joon, et al. "Deepsdf: Learning continuous signed distance functions for shape representation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Relation to the sequence modeling

Given that present ROM are more and more similar to the sequence modeling in lots of CS application, i.e. non-intrusive method, similar transformer network. I personally think it worth to think carefully about their relationship.

* Seq2Seq seems still not prevalent in scientific time series modeling.

* Stability and out-of-distribution issue