Supplementary Information

Weizhen Li[1,2,*], Qiang Fu[1,3], Yifan Hong[1], and Haojian Lu[2]

[1]Hangzhou Zhiyuan Research Institution, Hangzhou, Zhejiang, 310024, China
[2]College of Control Science and Engineering, Zhejiang University, 38 Zheda Road (Yuquan Campus), Hangzhou, Zhejiang, 310027, China
[3]Polytechnic Institute, Zhejiang University,269 Shixiang Road, Hangzhou, Zhejiang, 310015, China
[*]weizhen.li@zju.edu.cn

S1 Workflow of Sub-KNAFE

We present the complete, step-by-step workflow of the proposed subset regression with knowing the number of active features (sub-KNAFE) framework (taking back-KNAFE as an example, the most high-performing variant in our experiments). The schematic is shown in Fig. 1 of the main context.

1. Data Preprocessing: Given the observed state trajectory data $\mathbf{X}$ of the dynamical system, apply the automatic Savitzky-Golay filter[1,2] to smooth the data and compute the time derivatives $\dot{\mathbf{X}}$, to suppress noise and numerical differentiation errors.

2. Candidate Library Construction: Construct a unified polynomial candidate library $\mathbf{\Theta}(\mathbf{X})$ with a pre-specified maximum order, which contains all monomial terms up to the maximum order, with no terms pre-excluded.

3. Candidate Sub-Model Generation: Apply classical backward elimination to generate a complete sequence of candidate sub-models:

- Start with the full model containing all terms in the candidate library;
- At each step, eliminate the least significant term (with the highest p-value from t-test) and refit the model via least squares;
- Stop when the model is reduced to the null model, resulting in a sequence of sub-models ordered by the number of active features (from the full library size down to zero).

4. KNAFE-Based Model Selection: Using the prior known number of active features p, directly select the sub-model with exactly p active features from the pre-generated sequence of candidate sub-models.

5. Coefficient Estimation: Refit the selected sub-model via ordinary least squares to obtain the final coefficient estimates, and output the identified governing equation.

For for-KNAFE and bid-KNAFE, the workflow is identical except for the candidate sub-model generation step, which uses forward selection and bidirectional elimination respectively.

S2 The Power System Case

S2.1 Power Generator Rotor Dynamics

Power generator rotor dynamics can be described by the swing equation, a set of differential algebraic equations, defined as:

$$\begin{aligned} \dot{\boldsymbol{\delta}} &= \boldsymbol{\omega}, \\ M\dot{\boldsymbol{\omega}} &= \mathbf{P_{mesh}} - \mathbf{P_{elec}} - D\boldsymbol{\omega} + \boldsymbol{u}, \end{aligned} \tag{S1}$$

where vectors $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_r]^T$ and $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_r]^T$ correspond to the rotor angles and angular frequencies of all r generators. The equation parameters are defined as follows: $M = \text{diag}[M_1, \ldots, M_r]$ is a diagonal matrix containing the inertia constants for each generator, providing insight into their respective abilities to resist changes in rotational speed; $D = \text{diag}[D_1, \ldots, D_r]$ denotes the damping coefficients essential in offsetting power fluctuations and maintaining operational equilibrium; vectors $\mathbf{P_{mech}}$ and $\mathbf{P_{elec}}$ represent each generator's mechanical and electrical power. Furthermore, $\boldsymbol{u}$ signifies the external inputs that may include periodic forces exerted on the generator shafts, such as inappropriate parameters in the control loop from the turbine governor or exciter. In the study of forced oscillations, the external input $\boldsymbol{u}$ is particularly interesting as it may embody the periodic forcing factors that disrupt generator operations.Due to the periodic nature, the external input of the jth generator $\boldsymbol{u}_j$ can be transformed to the sum of sinusoidal components by the Fourier series:

$$\begin{aligned} u_j(t) &= \sum_{i=1}^{l} \Big(a_{ij} \sin\left(\omega_{F_{ij}} t\right) + b_{ij} \cos\left(\omega_{F_{ij}} t\right) \Big) \\ &= \sum_{i=1}^{l} \Big(\sqrt{a_{ij}^2 + b_{ij}^2} \sin(\omega_{F_{ij}} t + \varphi_{ij} \Big) \\ &= \sum_{i=1}^{l} \Big(\sqrt{\zeta_{ij}} \sin\left(\omega_{F_{ij}} t + \varphi_{ij}\right) \Big), \end{aligned} \tag{S2}$$

where $\omega_{F_{ij}}$ is the ith the dominant forced oscillation frequency for the jth generator, $\zeta_{ij} = \sqrt{a_{ij}^2 + b_{ij}^2}$ is the resultant amplitude, and $\varphi_{ij} = \arctan\left(\frac{b_{ij}}{a_{ij}}\right)$ is the phase angle for each harmonic component within the generator's input.

The transition from deterministic to stochastic modeling is essential in power system analysis because it enables the incorporation of inherent uncertainties in the system's operational dynamics. These uncertainties typically arise from fluctuations in load power, which are modeled as stochastic elements characterized by Gaussian noise. Understanding these stochastic variations is a key aspect of analyzing the behavior of the system, particularly at the generator's internal buses.To account for these load variations, the swing equations are expanded as

$$\begin{aligned} \dot{\boldsymbol{\delta}} &= \boldsymbol{\omega}, \\ M\dot{\boldsymbol{\omega}} &= \mathbf{P_{mech}} - \mathbf{P_{elec}} - D\boldsymbol{\omega} - E^2 G \Sigma \boldsymbol{\eta} + \boldsymbol{u}, \end{aligned} \tag{S3}$$

where $E = \text{diag}[E_1, \ldots, E_r]$ is the magnitude of internal electromotive force; $G = \text{diag}[G_{11}, \ldots, G_{rr}]$ denotes the conductance matrix; $\Sigma = \text{diag}[\sigma_{load_1}, \ldots, \sigma_{load_r}]$ the standard deviations of the load variations; $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_r]^T$, composed of standard Gaussian random variables, correlates to the noise in load power. This extension of the swing equation aims to encompass the stochastic nature of power systems, providing a more robust and realistic model that better reflects the unpredictable aspects of power system operation.

Under small perturbation, Eq. (S3) can be linearized around its steady-state operating point as below:

$$
\begin{aligned}
\begin{bmatrix} \Delta\dot{\boldsymbol{\delta}} \\ \Delta\dot{\boldsymbol{\omega}} \end{bmatrix} &= \begin{bmatrix} 0_{r\times r} & I_{r\times r} \\ -M^{-1}\frac{\partial \mathbf{P_{elec}}}{\partial \boldsymbol{\delta}} & -M^{-1}D \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\delta} \\ \Delta\boldsymbol{\omega} \end{bmatrix} \\
&+ \begin{bmatrix} \mathbf{0_{r\times 1}} \\ -M^{-1}E^2G\Sigma\boldsymbol{\eta} \end{bmatrix} + \begin{bmatrix} \mathbf{0_{r\times 1}} \\ \boldsymbol{u} \end{bmatrix}
\end{aligned} \tag{S4}
$$

By substituting $\boldsymbol{u}(t) = [u_1(t), \dots u_r(t)]^T$ in Eq. (S2) to the linearized swing equation, Eq. (S4), and simplifying notations, the power system model can be represented as:

$$
\begin{aligned}
\begin{bmatrix} \Delta\dot{\boldsymbol{\delta}} \\ \Delta\dot{\boldsymbol{\omega}} \end{bmatrix} &= \begin{bmatrix} 0_{r\times r} & I_{r\times r} \\ \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\delta} \\ \Delta\boldsymbol{\omega} \end{bmatrix} + \begin{bmatrix} \mathbf{0_{r\times 1}} \\ \boldsymbol{\beta}_0 \end{bmatrix} \\
&+ \begin{bmatrix} \mathbf{0} \\ \boldsymbol{a}_1 \end{bmatrix} \sin\left(\omega_{F_1} t\right) + \begin{bmatrix} \mathbf{0} \\ \boldsymbol{b}_1 \end{bmatrix} \cos\left(\omega_{F_1} t\right) + \cdots \\
&+ \begin{bmatrix} \mathbf{0} \\ \boldsymbol{a}_p \end{bmatrix} \sin\left(\omega_{F_p} t\right) + \begin{bmatrix} \mathbf{0} \\ \boldsymbol{b}_p \end{bmatrix} \cos\left(\omega_{F_p} t\right),
\end{aligned} \tag{S5}
$$

where $\boldsymbol{\beta}_0 = -M^{-1}E^2G\Sigma\boldsymbol{\eta}$ can be regarded as the intercept; $\boldsymbol{\beta}_1 = -M^{-1}\partial\mathbf{P_{elec}}/\partial\boldsymbol{\delta}$ and $\boldsymbol{\beta}_2 = -M^{-1}D$ are coefficients of $\Delta\boldsymbol{\delta}$ and $\Delta\boldsymbol{\omega}$; ω_{F_i}, $i = 1, \dots, p$ represents the dominant frequencies of forced oscillations affecting the power system, and a_{ij} and b_{ij} quantify the oscillation input magnitudes at the jth generator for each frequency ω_{Fi}.Thus, the response variable

$$
\begin{aligned}
\dot{\mathbf{X}} &= \begin{bmatrix} \Delta\dot{\boldsymbol{\delta}} & \Delta\dot{\boldsymbol{\omega}} \end{bmatrix} \\
&= \begin{bmatrix} \Delta\dot{\delta}_1(t_1) & \cdots & \Delta\dot{\omega}_r(t_1) \\ \vdots & \ddots & \vdots \\ \Delta\dot{\delta}_1(t_n) & \cdots & \Delta\dot{\omega}_r(t_n) \end{bmatrix}_{n\times 2r}
\end{aligned} \tag{S6}
$$

the candidate library

$$
\begin{aligned}
&\boldsymbol{\Theta}(\mathbf{X})_{n\times(1+2r+2p)} = \\
&\left[\begin{matrix} | & | & \cdots & | & | & \cdots & | \\ \mathbf{1} & \boldsymbol{\Delta\delta_1} & \cdots & \boldsymbol{\Delta\delta_r} & \boldsymbol{\Delta\omega_1} & \cdots & \boldsymbol{\Delta\omega_r} \\ | & | & \cdots & | & | & \cdots & | \end{matrix} \right. \\
&\left. \begin{matrix} \cdots & | & | & \cdots \\ \cdots & \boldsymbol{\sin(\omega_{F_i} t)} & \boldsymbol{\cos(\omega_{F_i} t)} & \cdots \\ \cdots & | & | & \cdots \end{matrix} \right],
\end{aligned} \tag{S7}
$$

and the coefficients matrix should be:

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_{\boldsymbol{\eta}} \\ \mathbf{B}_{\text{Jacobian}} \\ \mathbf{B}_{ab} \end{bmatrix}_{(1+2r+2p)\times 2r} \tag{S8}$$

$$\mathbf{B}_{\boldsymbol{\eta}} = \begin{bmatrix} 0 & \boldsymbol{\beta}_0 \end{bmatrix}_{1\times 2r} \tag{S9}$$

$$\mathbf{B}_{\text{Jacobian}} = \begin{bmatrix} 0 & \boldsymbol{\beta}_1 \\ I & \boldsymbol{\beta}_2 \end{bmatrix}_{2r\times 2r} \tag{S10}$$

$$\mathbf{B}_{ab} = \begin{bmatrix} 0 & \cdots & a_{11} & \cdots & a_{1r} \\ \vdots & \cdots & b_{11} & \cdots & b_{1r} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \vdots & \cdots & a_{p1} & \cdots & a_{pr} \\ 0 & \cdots & b_{p1} & \cdots & b_{pr} \end{bmatrix}_{2p\times 2r} . \tag{S11}$$

When the power system does not have forced oscillation, all elements in the matrix $\mathbf{B}_{ab}$ are zero. However, when the power system is under the influence of forced oscillations, the matrix $\mathbf{B}_{ab}$ is sparse, exhibiting significant non-zero values only at entries corresponding to the source generators at particular frequencies.

References

[1] Egan, K., Li, W. & Carvalho, R. Automatically discovering ordinary differential equations from data with sparse regression. *Commun. Phys.* **7**, 1–10 (2024).

[2] Li, W. & Carvalho, R. Automating the discovery of partial differential equations in dynamical systems. *Mach. Learn.: Sci. Technol.* **5**, 035046 (2024).