Computer Modeling in Engineering & Sciences
DOI:10.32604/cmes.2020.010061
Article

A Local Sparse Screening Identification Algorithm with Applications

Hao Li[1,2], Zhixia Wang[1,2] and Wei Wang[1,2,*]

[1]School of Mechanical Engineering, Tianjin University, Tianjin, 300350, China
[2]Tianjin Key Laboratory of Nonlinear Dynamics and Control, Tianjin, 300350, China
[*]Corresponding Author: Wei Wang. Email: wangweifrancis@tju.edu.cn
Received: 10 February 2020; Accepted: 11 May 2020

Abstract: Extracting nonlinear governing equations from noisy data is a central challenge in the analysis of complicated nonlinear behaviors. Despite researchers follow the sparse identification nonlinear dynamics algorithm (SINDy) rule to restore nonlinear equations, there also exist obstacles. One is the excessive dependence on empirical parameters, which increases the difficulty of data pre-processing. Another one is the coexistence of multiple coefficient vectors, which causes the optimal solution to be drowned in multiple solutions. The third one is the composition of basic function, which is exclusively applicable to specific equations. In this article, a local sparse screening identification algorithm (LSSI) is proposed to identify nonlinear systems. First, we present the *k*-neighbor parameter to replace all empirical parameters in data filtering. Second, we combine the mean error screening method with the SINDy algorithm to select the optimal one from multiple solutions. Third, the time variable *t* is introduced to expand the scope of the SINDy algorithm. Finally, the LSSI algorithm is applied to recover a classic ODE and a bi-stable energy harvester system. The results show that the new algorithm improves the ability of noise immunity and optimal parameters identification provides a desired foundation for nonlinear analyses.

Keywords: The *k*-neighbor parameter; sparse identification nonlinear dynamics algorithm; mean error screening method; the basic function; energy harvester

1 Introduction

For systems analysis, models are generally established using quantitative approaches. However, such quantitative methods are very effective for linear systems modelling not for nonlinear systems [1–3]. As most models are nonlinear, researchers have proposed various algorithms to recover nonlinear governing equations from time series. One of the most exciting modelling approaches is the sparse representation. Markus et al. [4] used sparse identification of nonlinear dynamics for rapid model recovery. Fahimeh et al. [5] identified nonlinear dynamical systems using the sparse recovery and dictionary learning. The continuous optimization of algorithm leads to the parameters of nonlinear system models being reconstructed using the neural network algorithm [6–8], genetic algorithm [9–12], and particle swarm algorithm [13–16]. Despite the promising performance of such algorithms, there are still some defects during the identification procedure. It is difficult for the genetic algorithm to solve the nonlinear constraint problems, which has a strong connection with initial population and empirical parameters [17].

Because of the random generation of the population, the results of the particle swarm algorithm deviate from the optimal solution [18]. Additionally, the neural network algorithm encounters difficulties in dealing with incomplete data. The most important aspect is that the developed strategy is often affected by the well-known overfitting problem [19].

Considering the limitations of general algorithms, Steven et al. [20] leveraged the fact that most physical systems have only a few relevant terms to define the dynamics, which made governing equations sparse in high-dimensional nonlinear function space, and proposed the sparse identification nonlinear dynamics algorithm (SINDy). The algorithm uses symbolic regression to determine the dynamics and conservation laws, and balances the complexity of the model (measured as the number of model terms) with the coherence of data. Moreover, it is a decision-making process for some empirical parameters based on the analysis and evaluation of expert knowledge in terms of dealing with noisy data. Simultaneously, as a result of the interference of nonlinear factors, there are redundant terms in the identified results, so the optimal solution cannot be automatically determined. According to Steven et al. [20], in successful examples, the choice of coordinates and initial conditions is somehow fortunate because they enable sparse representation.

Accordingly, in this article, we propose a local sparse screening identification algorithm (LSSI) that combines the local linear embedding (LLE) [21–23], the SINDy algorithm [24,25] and the mean error screening method (MES). The new algorithm replaces all empirical parameters and effectively avoids redundant terms in multiple identified solutions. In addition, the composition of the basic function is enhanced, and it is applied to a class of non-autonomous nonlinear systems. First, the LLE algorithm's k-neighbor parameter, which evolved from the hierarchical algorithm [26], is substituted for the selection of all empirical parameters, such as regularization parameters, step size, number of iterations, etc. This reduces the dependence on external experts and improves the precision of parameter selection. Meanwhile, the LLE algorithm has an inherent advantage in terms of data dimensionality reduction and noise filtering, which accordingly enhances noise robustness and accelerates high-dimensional system recovery from scratch. Second, the MES method automatically screens every possible solution to determine the optimal solution that improves the calculation efficiency. Third, the basic function introduces the time variable t to make it more complete and expands the scope of application of the SINDy algorithm.

The rest of the paper is organized as follows. Section 2 introduces the theory of the LSSI algorithm. Some important applications and comparisons between different approaches are presented in Section 3. The algorithm is applied to the data-driven modelling process of a classic ODE and a membrane type electromagnetic vibration energy harvester (EVEH), which shows promising results with respect to system identification, even starting from a strongly nonlinear and noisy reference dataset. The conclusions are drawn in Section 4.

2 The LSSI Algorithm

The aim of this algorithm is to solve the optimal sparse coefficients, subsequently governing equations are recovered. Fig. 1 demonstrates the LSSI algorithm, which contains three steps: data filtering, the SINDy algorithm and the MES method.

2.1 Data Filtering

The training dataset commonly includes random noise. The quality of noise filtering strongly depends on the setting of empirical parameters in testing process, such as regularization parameters, step size, number of iterations, etc. [27]. It means that any inappropriate choice of empirical parameter might finally lessen the robustness and performance of identification. The k-neighbor parameter in testing process is used as a

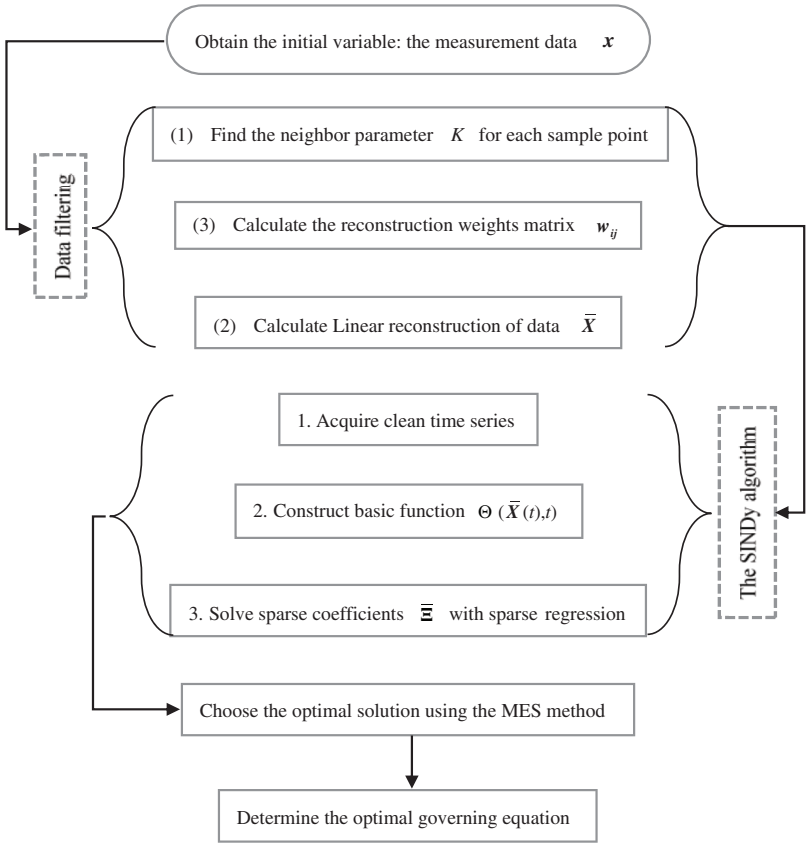


Figure 1: LSSI algorithm

substitute for all empirical parameters that are tough to choose, due to they strongly rely upon the analysis and evaluation of expert knowledge. The training dataset filtering process is divided into three steps.

(1) Find the neighbors for each sample point

In this study, the Euclidean distance measurement is used:

$$d(\boldsymbol{X},\boldsymbol{Y})=\|\boldsymbol{X}-\boldsymbol{Y}\|=\left[\sum_{i=1}^{n}(\boldsymbol{X_i}-\boldsymbol{Y_i})^2\right]^{1/2}. \tag{1}$$

(2) Calculate the reconstruction weight matrix w_{ij}

$$\min \varepsilon_i(W)=\sum_{i=1}^{n}\left\|\boldsymbol{X_i}-\sum_{j=1}^{K}\boldsymbol{w_{ij}}\boldsymbol{Y_j}\right\|_2^2=\sum_{i=1}^{n}\left[\sum_{j=1}^{K}\sum_{l=1}^{K}((\boldsymbol{X_i}-\boldsymbol{Y_j})^{\mathrm{T}}(\boldsymbol{X_i}-\boldsymbol{Y_l})\boldsymbol{w_{ij}}\boldsymbol{w_{il}})\right], \tag{2}$$

with the constraint condition

$$s.t\begin{cases}\sum_{j=1}^{n}\boldsymbol{w_{ij}}=1,\\ \boldsymbol{w_{ij}}=0,\boldsymbol{X_j}\notin N(\boldsymbol{X_i}),j=1,2,\cdots,K,\end{cases} \tag{3}$$

where $N(\boldsymbol{X_i})$ represents the neighbor points. When $\boldsymbol{X_j}$ is located in the range of $N(\boldsymbol{X_i})$, $\boldsymbol{w_{ij}}=1$; otherwise, $\boldsymbol{w_{ij}}=0$. Eq. (2) is calculated using the Lagrange coefficient.

Using the transition matrix $\boldsymbol{Z_i}=((\boldsymbol{X_i}-\boldsymbol{X_j})^{\mathrm{T}}(\boldsymbol{X_i}-\boldsymbol{X_j}))_{\boldsymbol{jl}} \in R^{K\times K}, (j,l=1,2,\cdots,K)$, Eq. (2) changes to

$$\min \varepsilon_i(W)=\sum_{i=1}^{n}\left\|\boldsymbol{X_i}-\sum_{j=1}^{K}\boldsymbol{w_{ij}}\boldsymbol{Y_j}\right\|_2^2=\sum_{i=1}^{n}\boldsymbol{w_i}^{\mathrm{T}}\boldsymbol{Z}_i\boldsymbol{w_i}. \tag{4}$$

Next, the weight matrix is

$$\frac{\partial L}{\partial w_i}=2\boldsymbol{Z_i}\boldsymbol{w_i}+\lambda\cdot 1_n=0 \Rightarrow \boldsymbol{w}_i=\frac{\boldsymbol{Z_i}^{-1}\cdot 1_K}{1_K^{\mathrm{T}}\boldsymbol{Z_i}^{-1}1_K}. \tag{5}$$

(3) Linear reconstruction of data using the *k*-neighbor parameter

Substituting $\boldsymbol{X_i}$ for the k-neighbor linear regression $\sum_{j=1}^{K}\boldsymbol{w_{ij}}\boldsymbol{X_j}$ of sample point $\boldsymbol{X_i}$ yields

$$\bar{\boldsymbol{X}}_{\boldsymbol{i}}=\sum_{j=1}^{K}\boldsymbol{w_{ij}}\boldsymbol{X_j} \Rightarrow \bar{\boldsymbol{X}}=\left[\bar{\boldsymbol{X}}_1 \quad \bar{\boldsymbol{X}}_2 \quad \cdots \quad \bar{\boldsymbol{X}}_n\right], \tag{6}$$

where $\bar{\boldsymbol{X}}$ is the filtered training sample, $\bar{\boldsymbol{X}}_{\boldsymbol{i}}(i\in(1,n))$ is a column vector of $\bar{\boldsymbol{X}}$.

2.2 The SINDy Algorithm

In 2016, Steven et al. [20] proposed the SINDy algorithm to identify nonlinear governing equations. That is

$$\dot{\boldsymbol{x}}(t)=\boldsymbol{f}(\boldsymbol{x}(\boldsymbol{t})), \tag{7}$$

In this work, we extend the sparse learning framework discussed the nonlinear non-autonomous system. Substituting $\bar{\boldsymbol{X}}$ into Eq. (7), we consider dynamics driven by a differential equation with external incentives,

$$\dot{\bar{\boldsymbol{X}}}(t)=\boldsymbol{f}(\bar{\boldsymbol{X}}(t)), \tag{8}$$

where the vector $\bar{\boldsymbol{X}}(t)\in R^n$ denotes the state of the system at time t, $\dot{\bar{\boldsymbol{X}}}(t)$ is the derivative of $\bar{\boldsymbol{X}}(t)$ and the function $\boldsymbol{f}(\bar{\boldsymbol{X}}(t))$ represents the dynamic constraints that define the equations of motion of the system, such as the momentum theorem.

Most physical systems have only a few nonlinear terms in their dynamics, which makes the right-hand side $\boldsymbol{f}(\bar{\boldsymbol{X}}(t))$ in Eq. (7) sparse in high-dimensional nonlinear function space. To search those remaining terms, we collect the time series $\bar{\boldsymbol{X}}(t)$ from the system structure and measures the derivative $\dot{\bar{\boldsymbol{X}}}(t)$ from $\bar{\boldsymbol{X}}(t)$. The dataset is sampled at several times $t_1,t_2,\cdots,t_m$ and the two matrices can be created in Eqs. (9) and (10),

$$\bar{\boldsymbol{X}}(t)=\begin{bmatrix}\bar{x}^T(t_1)\\ \bar{x}^T(t_2)\\ \vdots\\ \bar{x}^T(t_m)\end{bmatrix}=\overrightarrow{\begin{bmatrix}\bar{x}_1(t_1) & \bar{x}_2(t_1) & \cdots & \bar{x}_n(t_1)\\ \bar{x}_1(t_2) & \bar{x}_2(t_2) & \cdots & \bar{x}_n(t_2)\\ \vdots & \vdots & \ddots & \vdots\\ \bar{x}_1(t_m) & \bar{x}_2(t_m) & \cdots & \bar{x}_n(t_m)\end{bmatrix}}\ \text{state}\ \downarrow \text{time}, \tag{9}$$

$$\dot{\bar{\boldsymbol{X}}}(t)=\begin{bmatrix}\dot{\bar{x}}^T(t_1)\\ \dot{\bar{x}}^T(t_2)\\ \vdots\\ \dot{\bar{x}}^T(t_m)\end{bmatrix}=\overrightarrow{\begin{bmatrix}\dot{\bar{x}}_1(t_1) & \dot{\bar{x}}_2(t_1) & \cdots & \dot{\bar{x}}_n(t_1)\\ \dot{\bar{x}}_1(t_2) & \dot{\bar{x}}_2(t_2) & \cdots & \dot{\bar{x}}_n(t_2)\\ \vdots & \vdots & \ddots & \vdots\\ \dot{\bar{x}}_1(t_m) & \dot{\bar{x}}_2(t_m) & \cdots & \dot{\bar{x}}_n(t_m)\end{bmatrix}}\ \text{state}\ \downarrow \text{time}. \tag{10}$$

Depending on the above analysis, the sparse regression problem is

$$\dot{\bar{X}}(t) = \Theta(\bar{X}(t),t)\bar{\Xi}, \tag{11}$$

$$\Theta(\bar{X}(t),t)=\begin{bmatrix} | & | & | & & | & | & | & & | & | \\ 1 & \bar{X} & \bar{X}^{p_2} & \cdots & \bar{X}^{p_n} & \sin(t) & \cos(t) & \cdots & \sin(\omega t) & \cos(\omega t) \\ | & | & | & & | & | & | & & | & | \end{bmatrix}, \tag{12}$$

where $\Theta(\bar{X}(t),t)$ is the potential function of the columns $\bar{X}(t)$. It can be observed through simulation or measurement data according to the given initial conditions. Since the research object is a non-autonomous system, the improved basis function is shown in Eq. (12), which commonly consists of constants, polynomials, and trigonometric functions. $\sin(\omega t)$ and $\cos(\omega t)$ denote the external incentive, where ω is the excitation frequency. Each column of $\Theta(\bar{X}(t),t)$ is a candidate function for $\boldsymbol{f}(\bar{X}(t))$. The higher polynomials are denoted as $\bar{X}^{p_2}, \cdots, \bar{X}^{p_j}$, where $\bar{X}^{p_j}$ denotes the j^{th} nonlinearities in the state $\bar{X}(t)$, that is

$$\bar{X}^{p_j}=\begin{bmatrix} \bar{x}_1^j(t_1) & \bar{x}_2^j(t_1) & \cdots & \bar{x}_1^{j-1}(t_1)\bar{x}_n(t_1) & \cdots & \bar{x}_n^j(t_1) \\ \bar{x}_1^j(t_2) & \bar{x}_2^j(t_2) & \cdots & \bar{x}_1^{j-1}(t_2)\bar{x}_n(t_2) & \cdots & \bar{x}_n^j(t_2) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_1^j(t_m) & \bar{x}_2^j(t_m) & \cdots & \bar{x}_1^{j-1}(t_m)\bar{x}_n(t_m) & \cdots & \bar{x}_n^j(t_m) \end{bmatrix}. \tag{13}$$

The sparse regression problem is set up to determine the sparse coefficients $\bar{\Xi}=\begin{bmatrix}\bar{\xi}_1 & \bar{\xi}_2 & \cdots & \bar{\xi}_n\end{bmatrix}$, as described in Eq. (11). It starts with a least-squares solution for $\bar{\Xi}$ and then filters all coefficients that are smaller than the cut off values. That is, $\bar{\Xi}$ becomes the minimizer of

$$\tilde{\bar{\Xi}}=\arg\min_{\bar{\Xi}\in R}\left\|\dot{\bar{X}}(t)-\Theta(\bar{X}(t),t)\bar{\Xi}\right\|. \tag{14}$$

In general, the solution $\tilde{\bar{\Xi}}$ of Eq. (14) includes multiple solutions, as shown in Tab. 2, particularly for complex nonlinear systems. Consequently, the following MES method is used to determine the optimal one.

2.3 The MES Method

To solve that problem, the MES method is introduced to automatically select the optimal one among the identified results, which determines the minimum mean error using Pareto front analysis,

$$MES=\frac{1}{M}\sum\left(\dot{\bar{X}}(t)-\Theta(\bar{X}(t),t)\tilde{\bar{\Xi}}\right)^2, \tag{15}$$

where M is the number of the solution. The principle of MES method is that the minimum mean error corresponds to the optimal solution.

3 Application

The LSSI algorithm reduces the reliance on the selection of empirical parameters, automatically determines the optimal solution and expands the scope of adaptation. We verify the superiority of this new algorithm by modelling a classic ODE and a bi-stable EVEH.

3.1 Recovery of a Classic ODE Based on Numerical Simulation Data

The data we obtain from the physical experiments generally contains noise. Therefore, noise contained in a dataset should be considered to simulate a real-sense environment

$$\begin{cases} G^{\delta}=g+\delta C_0 e^{\delta}, \\ C_0=\dfrac{\|g\|_2}{\|e^{\delta}\|_2}, \end{cases} \tag{16}$$

where g is the original data, G^{δ} is noisy data, δ is the disturbance value, e^{δ} represents n random values ($n \in (0,1)$), and C_0 is a constant noise term.

The LSSI algorithm is expanded to consider a general model

$$\ddot{x}+\omega_0^2 x+\varepsilon((\alpha_1 x^2+\alpha_2 x^3)+\dot{x}(\beta_1+\beta_2 x^2))=\varepsilon F_0 \cos(\Omega_0 t), \tag{17}$$

which is available for a series of nonlinear systems [28,29]. Eq. (17) is a weakly nonlinear system with $\varepsilon \ll 1$; contrarily, it is a strongly nonlinear system. Additionally, we also assume $C_1=\omega_0$, $C_2=\beta_1$, $C_3=\alpha_1$, $C_4=\alpha_2$, $C_5=\beta_2$ and $C_6=F_0$.

The equation is given parameter values, as shown in Tab. 1. We set $x_0=2, \dot{x}_0=0, \delta=0.001$, and $\varepsilon=1$ as the initial values to numerically obtain the training dataset from Eq. (17), which includes a sequence of time states $\boldsymbol{x}$ and derivatives $\dot{\boldsymbol{x}}$, where $\dot{\boldsymbol{x}}$ is computed using cubic spline interpolation. The basic function is

$$\Theta(\boldsymbol{x}(t),\boldsymbol{t})=\begin{bmatrix} | & | & | & | & & | & | & & | \\ x(t) & \dot{x}(t) & x(t)^2 & x(t)\dot{x}(t) & \cdots & \dot{x}(t)^3 & \sin(t) & \cdots & \cos(3t) \\ | & | & | & | & & | & | & & | \end{bmatrix}, \tag{18}$$

which determines the equation by calculating the related sparse coefficients $\tilde{\bar{\Xi}}=\begin{bmatrix} \bar{\xi}_1 & \bar{\xi}_2 & \cdots & \bar{\xi}_n \end{bmatrix}$.

Table 1: Parameter values in Eq. (17)

ω_0	β_1	β_2	α_1	α_2	F_0	Ω_0
1	–1	2	–2	4	5	2

Table 2: Multiple solutions of the SINDy algorithm

$\tilde{\bar{\Xi}}$	1	x	$\dot{x}$	x^2	$x\dot{x}$	$\dot{x}^2$	x^3	$x^2\dot{x}$	$x\dot{x}^2$	$\dot{x}^3$	$\cos(2t)$
S_2	–0.3779	0	1.1573	2.0821	0	0	–4.0703	–2.0978	0	0	3.4616
S_3	–0.4550	0	1.1756	2.1504	0	0	–3.9797	–2.0112	0	0	3.2750
⋮						⋮					
S_5	2.5650	–6.5075	0	1.0892	0	0	–4.2380	–1.4701	0	0	14.5353
S_6	0	–1.0147	0.9929	2.0138	0	0	–4.0020	–1.9885	0	0	5.0195
S_7	0	–0.9523	0.9952	2.0155	0	0	–4.0097	–1.9918	0	0	4.9543
⋮						⋮					
S_{19}	0	–1.0978	0.9966	2.0313	0	0	–3.9207	–1.9974	0	0	4.9891
⋮						⋮					

$S_i(i=1,2,\cdots,n_s)$ denotes the solution of the identified results and n_s is the solution number.

Subsequently, we test the training dataset depending on the given parameter values. Fig. 2 demonstrates the procedure for successful identification from a given simulation dataset. Remarkably, it represents our innovation computing architecture that combines data filtering (LLE), sparse regression (SINDy) and optimal solution selection (MES).

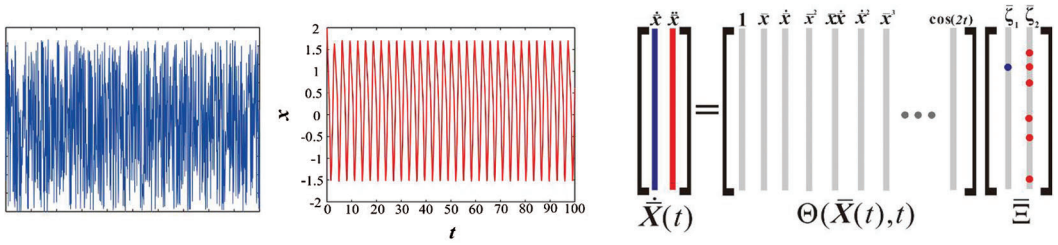


Figure 2: Strategy for the identification of Eq. (17) using the LSSI algorithm. (a) Noise level. (b) Noise-free time series curve. (c) The form of an ODE. The training dataset consists of (a) and (b). First, apply the hierarchical method to calculate the *k*-neighbor, as shown in Fig. 4, to complete the data filtering process. Second, determine the optimal solution S_6 from multiple solutions in Tab. 2. Finally, synthesize active terms in S_6, the optimal solution, into an ODE. The results are showed in Fig. 7

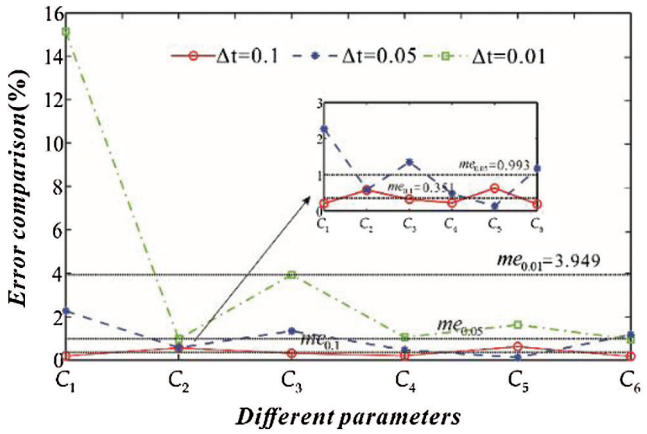


Figure 3: Identification results under different step sizes Δt in the filtering process. Each parameter C_i corresponds to the error in Eq. (17)

3.1.1 Replace Empirical Parameters Using a k-Neighbor

To demonstrate the effect of empirical parameters in filtering process, different step sizes are considered as a practical example, which directly affects the precision of governing equations. As shown in Fig. 3, contrary to our initial speculation, the larger the step size we increase, the more accurate the identification results are. Consequently, we replace empirical parameters to lessen uncertainty in data filtering.

The value of K relates to the global reconstruction error $\varepsilon(K)$. Figs. 4a and 4b indicate the obtained global *k*-neighbor. The error tolerance is limited to obtain a local *k*-neighbor, which is extracted from the global *k*-neighbor, as shown in Fig. 4c. The principle is that the minimum residual variance leads to the optimal values, where $K_{xopt} = 7$ and $K_{\dot{x}opt} = 8$. Subsequently, data filtering can be completed with K_{xopt} and $K_{\dot{x}opt}$ according to Eq. (6). Fig. 5 denotes the effect of the setting parameters on the identification precision of Eq. (17). In the noise level $\delta = 0.001$ and nonlinear disturbance $\varepsilon = 1$ case of, the SINDy algorithm fluctuates significantly, while the LSSI algorithm remains relatively stable. It is seen that the LSSI algorithm has high precision and gets rid of the dependence on traditional empirical parameters.

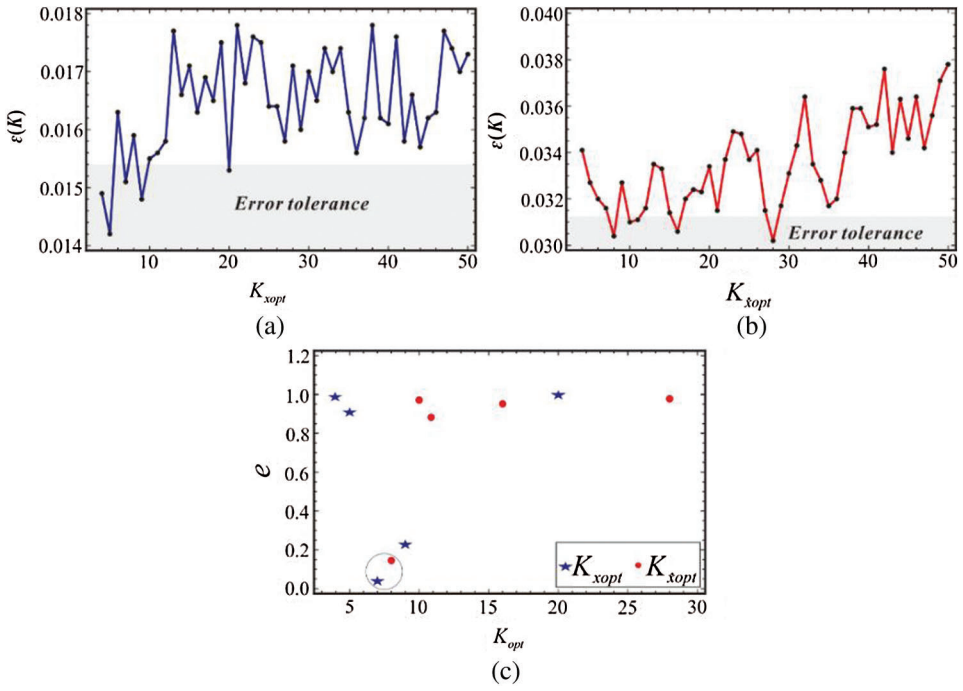


Figure 4: (a) and (b) Global reconstruction errors $\varepsilon(K)$ of $\boldsymbol{x}$ and $\dot{\boldsymbol{x}}$ for the k-neighbor, respectively. (c) Residual variance distribution of the optimal solutions K_{xopt} and $K_{\dot{x}opt}$. The shaded area represents the error tolerance, where the setting range considers sample numbers and nonlinear orders. According to the local k-neighbor, we obtain each residual variance value of K_{xopt} and $K_{\dot{x}opt}$

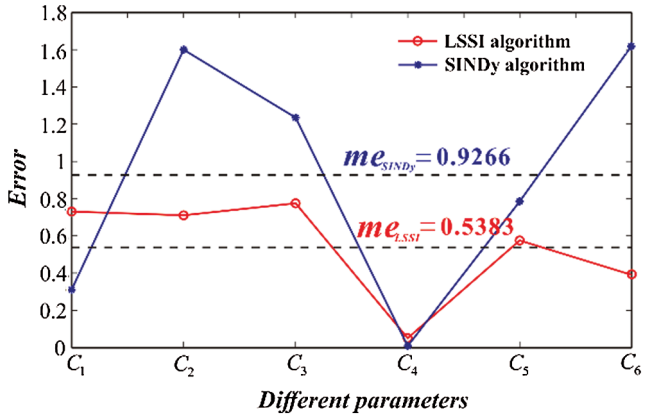


Figure 5: Calculation error under different parameters for the SINDy and LSSI algorithms. It describes the relationship between the error and each parameter value in Eq. (17)

3.1.2 Identify the Optimal Solution Using MES Method

In the SINDy algorithm, we obtain the sparse coefficients with multiple solutions. If the goal is to find the optimal solution that reliably represents the data among the large number of possibilities offered in the function, screening of $\tilde{\tilde{\Xi}}$ needs to be enforced and the process will be discussed below.

Multiple solutions with some redundant terms exist in Eq. (11), as shown in Tab. 2. A different solution S_i leads to a different model structure of the system. However as the model rises, sparse vectors $\tilde{\tilde{\Xi}}$ consequently

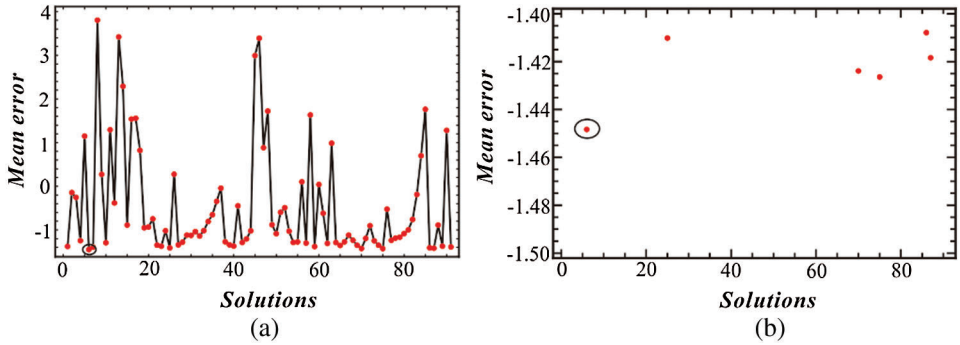


Figure 6: (a) Mean error distribution *vs.* the solutions in multiple solutions screening graph. (b) Close-up view of the mean error at crucial solutions

produce mean errors. Hence, the MES method is applied to select the optimal one in the mean error span, and it provides the minimum mean error that corresponds to the optimal solution S_6 in Fig. 6.

Comparison of the two algorithms denotes the LSSI algorithm whose advantage is demonstrated by its high computational precision in Tab. 3. The proposed algorithm with such sturdy noise resistance can determine the coefficients to be within 1% of the true values. We extract from Eq. (17) the time history curve that varies in the range of 5–12, as shown in Fig. 7. The local enlargement plot of the curve inside a period indicates that the LSSI algorithm is close to the original data, which similarly proves its high precision.

Table 3: Errors comparison between the SINDy and LSSI algorithms

Terms	Original coefficient	Identified results		Errors (%)	
		SINDy [20]	LSSI	SINDy [20]	LSSI
ω_0	1	1.0031	1.0073	0.3100	0.7300
β_1	–1	–0.9840	–0.9929	1.6000	0.7100
α_1	–2	–2.0247	–2.0155	1.2350	0.7750
α_2	4	3.9997	4.0020	0.0075	0.0500
β_2	2	1.9843	1.9885	0.7850	0.5750
F_0	5	4.9189	5.0195	1.6220	0.3900
me_{name}				0.9266	0.5383

Tab. 1 Section 3 p. 6 in [ω_0], [β_1], [α_1], [α_2], [β_2], [F_0].

3.1.3 Analyse Noise Level and Nonlinear Perturbation

Considering the noise level δ and perturbation strength ε, the standard for judging the visual quality is presented by the mean error, as shown in Fig. 8a. Clearly, the mean error increases with the growth of δ and ε. In Fig. 8b, when the nonlinear perturbation $\varepsilon = 1$, noise level of different orders of magnitude is applied to the training samples, that is 0.001, 0.01, 0.1. In particular, the mean error of the SINDy algorithm will reach approximately 20% at $\delta = 0.1$. The training samples are seriously polluted by noise, which makes it difficult to recover governing equations in testing process. However, in the mentioned above case, the LSSI algorithm has the potential to suppress noise and keeps the error precision around 12%. Note that it is more suitable for noisy data model recovery. Tab. 4 shows the influence of different nonlinear perturbations on

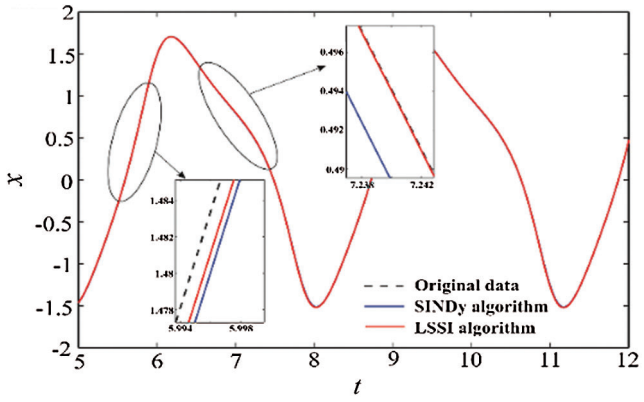


Figure 7: Contrast of the time history curve at $\Omega_0 = 2$ with an acceleration level of $F_0 = 5$ for the two algorithms

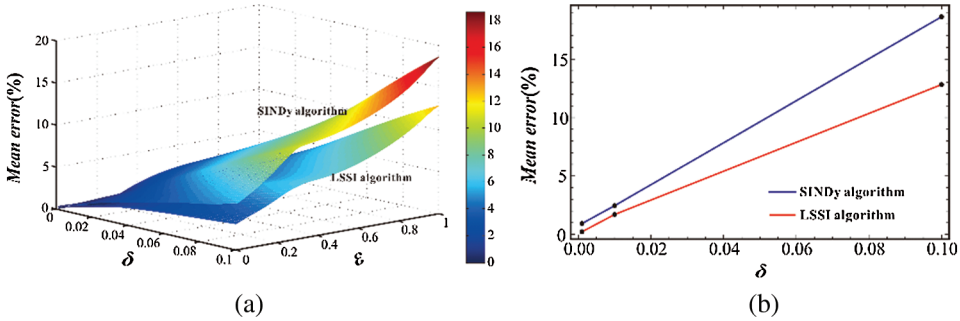


Figure 8: (a) Variation trend of the mean error under different disturbance δ and nonlinear strength ε for the SINDy and LSSI algorithms. (b) Relationship between the mean error and noise level at $\varepsilon = 1$

Eq. (17) under the same noise level. The new algorithm is approximately controlled within 1%, even with the strong perturbation $\varepsilon = 1$, so it is applicable to both strongly and weakly nonlinear systems.

3.2 Recovery of a Bi-Stable EVEH Based on Experimental Data

For the experimental dataset, we consider a bi-stable harvester. The model of this component is shown in Fig. 9. The working principle is that the concentric permanent magnet driven by the membrane vibrates forwards and backwards in the cavity wall, which changes the magnetic flux of the coil windings with an iron core around the front and back end covers to generate inductive electromotive force. The distance between the magnet and core influences the equilibrium position of the system. When the iron core is adjusted within a certain range, the EVEH performs in the bi-stable oscillation stage.

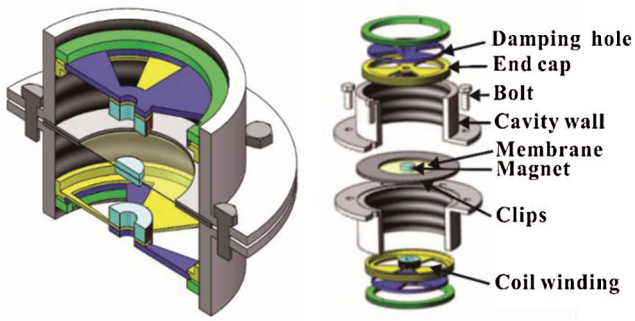
3.2.1 Theoretical Analysis of the EVEH

Given the physical parameters in Tab. 5, a theoretical equation can be established. In detail, governing equations describe the forced lateral axisymmetric vibration of a circular membrane with a centre magnet. Assuming that the rigid magnet sustains a front-back symmetric transverse vibration, an axial force is exerted as the boundary condition. Regarding the eigenfunction and the boundary conditions, the differential eigenvalue equations can be obtained. Discretising the partial differential equations obtains the ODEs. If no air resistance or random noise effects occur in the EVEH, the equation is given by [30].

Table 4: Errors of the LSSI algorithm under different nonlinear perturbations

Terms	Original coefficient	Identified results		Errors (%)	
		$\varepsilon = 0.3$	$\varepsilon = 1$	$\varepsilon = 0.3$	$\varepsilon = 1$
ω_0	1	0.9955	1.0073	0.4500	0.7300
β_1	–1	–1.0027	–0.9929	0.2700	0.7100
α_1	–2	–2.0003	–2.0155	0.0150	0.7750
α_2	4	4.0063	4.0020	0.1575	0.0500
β_2	2	2.0063	1.9885	0.3150	0.5750
F_0	5	4.9934	5.0195	0.1320	0.3900
me_{LSSI}				0. 2233	0.5383

Tab. 1 Section 3 p. 6 in [ω_0], [β_1], [α_1], [α_2], [β_2], [F_0].



(a) (b)

Figure 9: Structure of a membrane type bi-stable EVEH. (a) Structural profile. (b) Structural assembly drawing

Table 5: Physical parameters of the EVEH [30]

Parameters	Values
Radius of center magnet	7.5 mm
Radius of membrane	60 mm
Thickness of membrane	0.05 mm
Mass of center magnet	10.6×10^{-3} kg
Density of membrane	1420 kg/m^3
Elasticity modulus	90 MPa
Poisson's ratio	0.3
Tension of membrane	1 N/m
Linear stiffness coefficient	2.700 N/mm
Nonlinear stiffness coefficient	8.432×10^{-3} N/mm^3
Damping ratio	0.011
Resonant frequency (ω_1)	2.92 Hz

$$\ddot{q}+B\dot{q}+Dq+Hq^3=F\cos(\Omega t), \tag{19}$$

where B is the damping coefficient; D and H represent the squared term of the linear frequency and the cubic term coefficient, respectively; and $F\cos(\Omega t)$ is the external excitation, where Ω and F are the frequency and acceleration.

The detailed theoretical analysis of Eq. (19) is derived from [30]. We quote the theoretical results to be compared with experimental measurements and the LSSI algorithm. The following sections describe the procedure.

3.2.2 Experimental Analysis of the EVEH

The layout of the testing system is shown in Fig. 10. The EVEH is implemented using a Shaker (APS 400) that drives a signal generator (Agilent 33250A) with different frequencies and amplitudes. During the experiments, a data acquisition device (B & K3039) is used to record the vibration acceleration, displacement response and voltage outputs. Based on the displacement response, we apply the LSSI algorithm to recover governing equations. The main process is shown in Fig. 11.

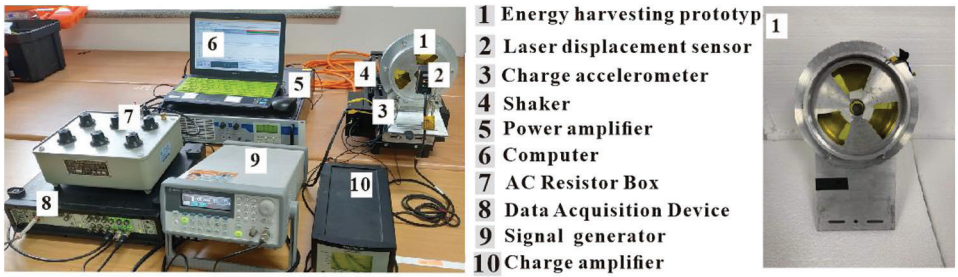


Figure 10: Experimental device schematic

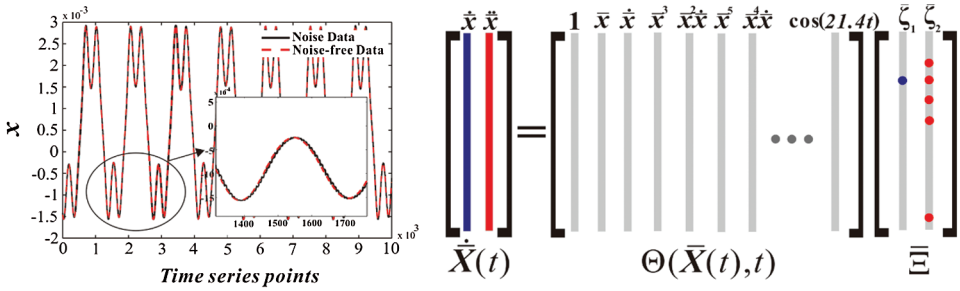


Figure 11: Steps in the LSSI algorithm for experimental data identification. (a) Time series curve. (b) The form of an ODE. First, data is collected as the time series curve from EVEH. Second, the noise-free data sample at an acceleration level of 0.98 m/s^2 is obtained by filtering the displacement signal, which is extracted from sweep signals. Third, the basis function is constructed from input sample data to solve $\bar{\Xi}$. Finally, the MES method selects the optimal solution S_{17} in Tab. 6. Active terms in S_{17} are synthesised into an ODE

In the experimental test, the partial displacement signal which contains noise level is extracted from the time series curve according to the sampling frequency. By analyzing and observing the frequency, the experimental measurement produces a resonant frequency with a value that reaches approximately 3.41 Hz. Fig. 11a shows the displacement signal with burrs smoothed to obtain the noise-free time series

during the data filtering process. Based on noise-free training dataset, the LSSI algorithm to recover governing equations, that is, the active terms (the optimal solution) are synthesizes into an ODE, as shown in Fig. 11b.

During the equation reconstruction process, the sparse coefficients $\tilde{\tilde{\Xi}}$ with multiple solutions exist in the EVEH system, as shown in Tab. 6. Subsequently, the MES method is applied to select the optimal solution. According to the principle of the minimum mean error, S_{17} is the optimal one, as shown in Fig. 12. Generalising the LSSI algorithm makes it possible to recover governing equations with different order nonlinearities. The resonant frequency is obtained from the identified results in Tab. 7, which is compared with theory and experiment in the following sections.

Table 6: Multiple solutions of the experimental dataset

$\tilde{\tilde{\Xi}}$	1	x	$\dot{x}$	x^3	$x^2\dot{x}$	$\cos(21.4t)$
S_1	–0.7702	-1.069×10^3	–15.8780	9.6068×10^7	5.8747×10^7	1.0458
⋮				⋮		
S_{17}	0	411.4801	–6.4302	-9.0418×10^7	-9.3898×10^6	0.3983
⋮				⋮		
S_{95}	0	524.2199	3.6873	-3.6366×10^8	-9.0408×10^6	0.2940

$S_i (i = 1, 2, \cdots, n_s)$ denotes the solution of the identified results and n_s is the solution number.

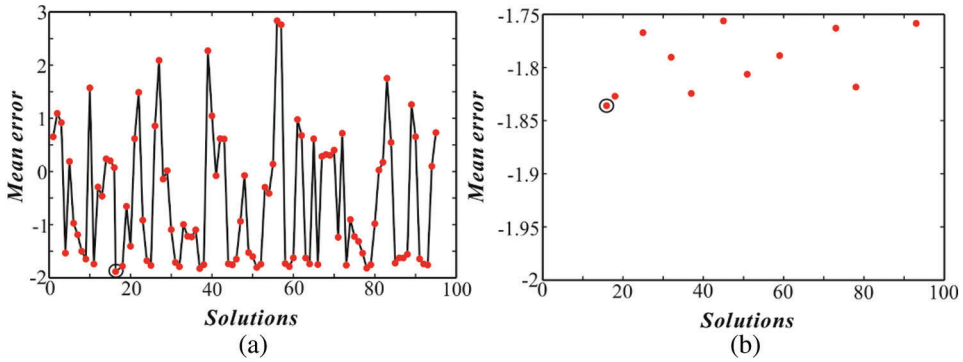


Figure 12: (a) Distribution plot of multiple solutions screening. (b) Close-up view of the mean error at crucial solutions

3.2.3 Comparative Analysis in Theory, Experiment and LSSI Identification

For governing equations with different order nonlinearities, the identified resonant frequency fluctuates in the range of 3–3.5 Hz, as shown in Fig. 13. It demonstrates that the equation with 9th-order nonlinear components is close to the original system, which can be regarded as the research foundation for future nonlinear analysis.

The discrepancy of the initial theoretical model in Eq. (19) originates from an insufficient consideration of complicated nonlinear factors in the membrane vibration, in addition to unavoidable measurement errors. According to Williams et al. [31,32], the energy harvester may be further constrained by unwanted damping owing to undesirable effects, such as air resistance. It is note that the deviations are hardness to be effectively exhibited in theoretical modelling process. Hence, data-driven modelling has become an inevitable choice to establish and improve governing equations of nonlinear vibration systems both numerically and experimentally.

Table 7: Multiple solutions of the experimental dataset

Terms	Order			
	O_3	O_5	O_7	O_9
1	0	0	0	0
x	411.4801	355.3357	374.7547	467.8553
$\dot{x}$	–6.4302	–11.9491	–3.0734	–7.6849
x^3	-9.0418×10^7	-4.0039×10^8	-9.5907×10^8	-1.2216×10^9
$x^2\dot{x}$	-9.3898×10^6	2.4695×10^7	-1.524×10^7	6.7945×10^7
x^5	0	6.5564×10^{13}	4.0994×10^{14}	7.5252×10^{14}
$x^4\dot{x}$	0	-1.2787×10^{12}	2.5527×10^{13}	-6.7882×10^{13}
x^7	0	0	-3.4641×10^{19}	-1.5368×10^{20}
$x^6\dot{x}$	0	0	-4.5204×10^{18}	1.9544×10^{19}
x^9	0	0	0	1.0016×10^{25}
$x^8\dot{x}$	0	0	0	-1.6465×10^{24}
$\cos(21.4t)$	0.3983	0.1126	0.2289	0.1571

$O_i(i = 3, 5, 7, 9)$ represents the highest order of nonlinear terms.

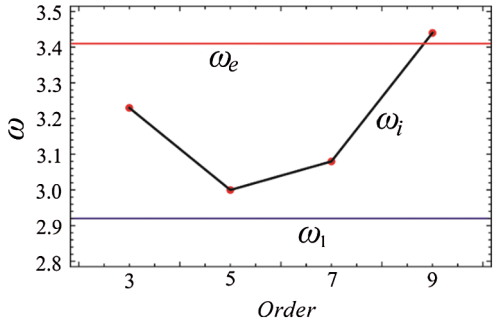


Figure 13: Resonant frequency comparison with the experimental measurement ($\omega_e = 3.41$ Hz), theoretical calculation ($\omega_1 = 2.92$ Hz), and identified results ($\omega_i(i = 3, 5, 7, 9)$). i denotes the order of the nonlinear term

Meanwhile, any infinite-dimensional continuous system expressed as partial differential equations can be discretized and expressed as a finite-dimensional matrix equation. Such a finite-dimensional system can often be expressed as a SDOF system using orthogonal transformations. Thus, the main concerned SDOF system in this article provides a good start point to the data-driven modelling process, based on which future research on more complicated systems can be built upon.

4 Conclusions

The LSSI algorithm displays high precision in recovering governing equations from experimental and numerical data. For the new algorithm, the k-neighbor parameter is substituted for all empirical parameters in data filtering to reduce the reliance on 'fortunate choice'. It also applies the MES method to solve the multi-solution problem in the sparse recovery procedure. The basic function has been enhanced to extend the scope of the SINDy algorithm. The innovations have shown promising results in terms of noise immunity and hidden nonlinear factor mining starting from a strongly nonlinear and noisy reference dataset.

The efficiency of the LSSI algorithm is verified in two stages. First, the identified equations are compared with the original SINDy algorithm in a classic ODE. Second, the theoretical model of an EVEH system is considered and then compensated necessary nonlinear components to the mechanical model to simulate the experimental dataset. The results show that there are promising potential applications in data-driven modelling process that arises across the physical, engineering, and biological sciences.

However, other multiscale systems or high-dimensional datasets that involve more complicated nonlinear terms may be encountered in practice. Therefore, how to ensure the physical meaning of nonlinear terms in the identification process is considered as an open question, which will be explored in future work.

Funding Statement: The work was supported by the National Science Foundation of China (grant nos. 11772218 and 11872044), China-UK NSFC-RS Joint Project (grant nos. 11911530177 in China and IE181496 in the UK), Tianjin Research Program of Application Foundation and Advanced Technology (grant no. 17JCYBJC18900).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Duan, C. C., Zhang, Y. F., Feng, C. (2018). Parameter identification of nonlinear system model based on DEAFCR algorithm. *Industrial Instrumentation & Auto, 6,* 3–6.
2. Liu, C. S., Kuo, C. L., Chang, J. R. (2020). Solving the optimal control problems of nonlinear duffing oscillators by using an iterative shape functions method. *Computer Modeling in Engineering & Sciences, 122(1),* 33–48. DOI 10.32604/cmes.2020.08490.
3. Elettreby, M. F., Nabil, T., Khawagi, A. (2020). Stability and bifurcation analysis of a discrete predator-prey model with mixed holling interaction. *Computer Modeling in Engineering & Sciences, 122(3),* 907–921. DOI 10.32604/cmes.2020.08664.
4. Quade, M., Abel, M., Kutz, J. N., Brunton, S. L. (2018). Sparse identification of nonlinear dynamics for rapid model recovery. *Chaos, 28(6),* 063116. DOI 10.1063/1.5027470.
5. Fahimeh, N., Aboozar, G., Sajad, J., Hasemi, G. (2019). Sparse recovery and dictionary learning to identify the nonlinear dynamical systems: one step toward finding bifurcation points in real systems. *International Journal of Bifurcation and Chaos, 29(3),* 1950030. DOI 10.1142/S0218127419500305.
6. Liao, J. L., Yin, F., Luo, Z. H., Chen, B., Sheng, D. R., Yu, Z. T. (2018). The parameter identification algorithm of steam turbine nonlinear servo system based on artificial neural network. *Journal of the Brazilian Society of Mechanical Science and Engineering, 40(165),* 165–175. DOI 10.1007/s40430-018-1086-8.
7. Chen, X., Kopsaftopoulos, F., Wu, Q., Ren, H., Chang, F. K. (2019). A self-adaptive 1D convolutional neural network for flight-state identification. *Sensors, 19(2),* 275. DOI 10.3390/s19020275.
8. Abrahim, A. R., Rahim, M. S., Su, L. (2019). Splicing image forgery identification based on artificial neural network approach and texture features. *Cluster Computing, 22(S1),* 647–660. DOI 10.1007/s10586-017-1668-8.
9. Sangdani, M. H., Tavakolpour-Saleh, A. R. (2018). Parameters identification of an experimental vision-based target tracker robot using genetic algorithm. *International Journal of Engineering, 31(3),* 480–486.
10. Ivanov, D. V., Engelgardt, V. V., Sandler, I. L. (2018). Genetic algorithm of structural and parametric identification of gegenbauer autoregressive with noise on output. *Procedia Computer Science, 131,* 619–625. DOI 10.1016/j.procs.2018.04.304.
11. Whelan, M., Zamudio, N. S., Kernicky, T. (2018). Structural identification of a tied arch bridge using parallel genetic algorithms and ambient vibration monitoring with a wireless sensor network. *Journal of Civil Structural Health Monitoring, 8(2),* 315–330. DOI 10.1007/s13349-017-0266-z.
12. Bartkowski, P., Zalewski, R., Chodkiewicz, P. (2019). Parameter identification of Bouc-Wen model for vacuum packed particles based on genetic algorithm. *Archives of Civil and Mechanical Engineering, 19(2),* 322–333. DOI 10.1016/j.acme.2018.11.002.

13. Khoja, I., Ladhari, T., Sakly, A., M'sahli, F. (2018). Parameter identification of an activated sludge wastewater treatment process based on particle swarm optimization algorithm. *Mathematical Problems in Engineering, 2018,* 1–11.
14. Wang, L., Liu, Y. Q. (2018). Application of simulated annealing particle swarm optimization based on correlation in parameter identification of induction motor. *Mathematical Problems in Engineering, 2018(6),* 1–9. DOI 10.1155/2018/1869232.
15. Chun, S., Kim, T. (2018). Simultaneous identification of model structure and the associated parameters for linear systems based on particle swarm optimization. *Complexity, 2018(10),* 1–17. DOI 10.1155/2018/2713684.
16. Damiani, L., Diaz, A. I., Iparraguirre, J., Blanco, A. M. (2020). Accelerated particle swarm optimization with explicit consideration of model constraints. *Cluster Computing, 23(1),* 149–164. DOI 10.1007/s10586-019-02933-1.
17. Wang, P. C., Chen, J., Wang, H. Q. (2019). Walking load identification based on two-stage genetic algorithm. *Journal of Vibration and Shock, 38(19),* 64–69.
18. Huang, M. S., Lei, Y. Z., Cheng, S. X. (2019). Damage identification of bridge structure considering temperature variations based on particle swarm optimization—cuckoo search algorithm. *Advances in Structural Engineering, 22(15),* 3262–3276. DOI 10.1177/1369433219861728.
19. Kapanova, K. G., Dimov, I., Sellier, J. M. (2018). A genetic approach to automatic neural network architecture optimization. *Neural Computing and Application, 29(5),* 1481–1492. DOI 10.1007/s00521-016-2510-6.
20. Brunton, S. L., Proctor, J. L., Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of National Academy of Sciences of the United States of America, 113(15),* 3932–3937. DOI 10.1073/pnas.1517384113.
21. Sun, L., Wang, W., Xu, J. C., Zhang, S. G. (2019). Improved LLE and neighborhood rough sets-based gene selection using Lebesgue measure for cancer classification on gene expression data. *Journal of Intelligent and Fuzzy Systems, 37(4),* 5731–5742. DOI 10.3233/JIFS-181904.
22. Roweis, S. T., Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290 (5500),* 2323–2326. DOI 10.1126/science.290.5500.2323.
23. Heidari, M., Moattar, M., Mohammad, H. (2019). Discriminative geodesic Gaussian process latent variable model for structure preserving dimension reduction in clustering and classification problems. *Neural Computing and Applications, 31(8),* 3265–3278. DOI 10.1007/s00521-017-3273-4.
24. Loiseau, J. C., Noack, B. R., Brunton, S. L. (2018). Sparse reduced-order modelling: sensor-based dynamics to full-state estimation. *Journal of Fluid Mechanics, 844,* 459–490. DOI 10.1017/jfm.2018.147.
25. Goharoodi, S. K., Dekemele, K., Dupre, L., Loccufier, M., Crevecoeur, G. (2018). Sparse identification of nonlinear duffing oscillator from measurement data. *IFAC-Papers Online, 51(33),* 162–167. DOI 10.1016/j.ifacol.2018.12.111.
26. Kouropteva, O., Okun, O., Pietikainen, M. (2002). Selection of the parameter value for the locally linear embedding algorithm. *Fuzzy Systems and Knowledge Discovery, 2,* 359–363.
27. Mahmoudi, G., Fouladi, M. R., Ay, M. R., Rahmim, A., Ghadiri, H. (2019). Sparse-view statistical image reconstruction with improved total variation regularization for X-ray micro-CT imaging. *Journal of Instrumentation, 14(8),* P08023. DOI 10.1088/1748-0221/14/08/P08023.
28. Xia, P., Xu, H., Lei, M. H., Ma, Z. C. (2018). An improved stochastic resonance algorithm with arbitrary stable-state matching in underdamped nonlinear systems with a periodic potential for incipient bearing fault diagnosis. *Measurement Science and Technology, 29(8),* 085002. DOI 10.1088/1361-6501/aac733.
29. Wang, W., Cao, J. Y., Mallick, D., Roy, S. (2018). Comparison of harmonic balance and multi-scale algorithm in characterizing the response of mono-stable energy harvesters. *Mechanical Systems and Signal Processing, 108,* 252–261. DOI 10.1016/j.ymssp.2018.02.035.
30. Zhang, B., Wang, W., Zhang, Q. C. (2019). Dynamic modelling and structural optimization of a bistable electromagnetic vibration energy harvester. *Energies, 12(12),* 2410. DOI 10.3390/en12122410.
31. Williams, C. B., Shearwood, C., Harradine, M. A., Mellor, P. H., Yates, R. B. (2001). Development of an electromagnetic micro-generator. *IEE Proceedings–Circuits, Devices and Systems, 148(6),* 337–342. DOI 10.1049/ip-cds:20010525.
32. Williams, C. B., Yates, R. B. (1996). Analysis of a micro-electric generator for microsystems. *Sensors and Actuators A: Physical, 52(1–3),* 8–11. DOI 10.1016/0924-4247(96)80118-X.

Appendix A

Nomenclature

$\bar{\boldsymbol{X}}(t)$	Noisy data
$\boldsymbol{f}(\bar{\boldsymbol{X}}(t))$	Nonlinear system
$\Theta(\bar{\boldsymbol{X}}(t),t)$	Basic function
$\bar{\boldsymbol{X}}^{p_j}(j=2,3,\cdots)$	Higher polynomials
$\bar{\boldsymbol{\Xi}}$	Coefficient terms
$\bar{\boldsymbol{\xi}}_i(i=1,2,\cdots)$	Nonlinear coefficients
Δt	Step size
$\boldsymbol{w_{ij}}$ (or $\boldsymbol{w_{il}}$)	Weight matrix
$\varepsilon_i(W)$	The minimum reconstruction errors
$d(\boldsymbol{X},\boldsymbol{Y})$	Euclidean distance
$\boldsymbol{Z_i}$	Transition matrix
K	The number of neighbors
ε	Nonlinear strength
ω_0	Natural frequency
$\alpha_i(i=1,2)$	Nonlinear terms coefficients
$\beta_j(j=1,2)$	Damping coefficients
Ω_0	External excitation frequency
F_0	External excitation amplitude
$C_t(t=1,\cdots,6)$	Nonlinear system terms
$me_{name}(name=LSSI,SINDy)$	Mean error
g	Original data
G^{δ}	Noisy data
δ	Disturbance value
e^{δ} $(n\in(0,1))$	Random values
C_0	Constant noise term
e	Residual variance
$\varepsilon(K)$	Local minimum values
W	Squared term of frequency
ω_e	Experimentally identified resonant frequency

(Continued)

(continued).	
ω_1	Theoretically calculated resonance frequency
$\omega_i (i = 3, 5, 7, 9)$	Identified resonance frequency
$S_i (i = 1, \cdots, n)$	The solution of the identified results
$O_i (i = 3, 5, 7, 9)$	The highest order of nonlinear terms