Chaos

How entropic regression beats the outliers problem in nonlinear system identification EP

Cite as: Chaos **30**, 013107 (2020); https://doi.org/10.1063/1.5133386
Submitted: 24 October 2019 . Accepted: 09 December 2019 . Published Online: 06 January 2020

Abd AlRahman R. AlMomani, Jie Sun, and Erik Bollt

COLLECTIONS

Note: This paper is part of the Focus Issue, "When Machine Learning Meets Complex Systems: Networks, Chaos and Nonlinear Dynamics."

EP This paper was selected as an Editor's Pick

View Online Export Citation CrossMark

Scilight
Highlights of the best new research in the physical sciences
LEARN MORE!
AIP Publishing

AIP Publishing

© 2020 Author(s).

How entropic regression beats the outliers problem in nonlinear system identification

Cite as: Chaos **30**, 013107 (2020); doi: 10.1063/1.5133386
Submitted: 24 October 2019 · Accepted: 9 December 2019 ·
Published Online: 6 January 2020

View Online Export Citation CrossMark

Abd AlRahman R. AlMomani,[1,2,a)] Jie Sun,[3,b)] and Erik Bollt[1,2,c)]

AFFILIATIONS

[1] Electrical and Computer Engineering, Clarkson University, Potsdam, New York 13699, USA
[2] Clarkson Center for Complex Systems Science (C^3S^2), Potsdam, New York 13699, USA
[3] Theory Lab, Hong Kong Research Centre of Huawei Tech, Hong Kong 852, China

Note: This paper is part of the Focus Issue, "When Machine Learning Meets Complex Systems: Networks, Chaos and Nonlinear Dynamics."

[a)] **Electronic mail:** aaalmoma@clarkson.edu
[b)] **Electronic mail:** sunj@clarkson.edu
[c)] **Electronic mail:** bolltem@clarkson.edu

ABSTRACT

In this work, we developed a nonlinear System Identification (SID) method that we called Entropic Regression. Our method adopts an information-theoretic measure for the data-driven discovery of the underlying dynamics. Our method shows robustness toward noise and outliers, and it outperforms many of the current state-of-the-art methods. Moreover, the method of Entropic Regression overcomes many of the major limitations of the current methods such as sloppy parameters, diverse scale, and SID in high-dimensional systems such as complex networks. The use of information-theoretic measures in entropic regression has unique advantages, due to the Asymptotic Equipartition Property of probability distributions, that outliers and other low-occurrence events are conveniently and intrinsically de-emphasized as not-typical, by definition. We provide a numerical comparison with the current state-of-the-art methods in sparse regression, and we apply the methods to different chaotic systems such as the Lorenz System, the Kuramoto-Sivashinsky equations, and the Double-Well Potential.

Published under license by AIP Publishing. https://doi.org/10.1063/1.5133386

System identification (SID) is a central concept in science and engineering applications whereby a general model form is assumed, but active terms and parameters must be inferred from observations. Most methods for SID rely on optimizing some metric-based cost function that describes how a model fits observational data. A commonly used cost function employs a Euclidean metric and leads to a least squares (LS) estimate, whereas recently it has become popular to also account for model sparsity such as in compressed sensing (CS) and Lasso. While the effectiveness of these methods has been demonstrated in previous studies, including in cases where outliers exist in sparse samples, SID remains particularly difficult under more realistic scenarios where each observation is subject to non-negligible noise and is sometimes even contaminated by large noise outliers. Here we report that existing sparsity-focused methods such as compressive sensing, when applied in such scenarios, can result in "oversparse" solutions that are brittle to outliers. In fact, metric-based methods are prone to outliers because outliers by nature have a disproportionately large influence. To mitigate such issues of large noise and outliers, we develop an entropic regression approach for nonlinear SID, whereby true model structures are identified based on an information-theoretic criterion describing relevance in terms of reducing information flow uncertainty vs not necessarily (just) sparsity. The use of information-theoretic measures in entropic regression has unique advantages due to the asymptotic equipartition property (AEP) of probability distributions, that outliers and other low-occurrence events are conveniently and intrinsically de-emphasized as not-typical by definition.

A basic and fundamental problem in science and engineering is to collect data as observations from an experiment and then to attempt to explain the experiment by summarizing data in terms of a model. When dealing with a dynamical process, a common scenario is to describe the underlying process as a dynamical system, which may be in the form of a differential equation (DE). Traditionally, this means "understanding the underlying physics" in a manner that allows one to write a DE from first principles, including those terms that capture the delicate but important (physical) effects. The validation of the model may come from comparing outputs from the model to those from experiments, where outputs are typically represented as multivariate time series. Building a DE model based on fundamental laws and principles requires strong assumptions, which might be evaluated by how the model fits data. Weigend and Gershenfeld made a distinction between weak modeling (data rich and theory poor) and strong modeling (data poor and theory rich), and suggest that it is related to "…the distinction between memorization and generalization.…"[1]

The problem of learning a (dynamical) system from observational data is known as *system identification* (SID) and oftentimes involves the underlying assumption that the *structural* form of the DE is known (which kinds of terms to include in the functional description of the equation), but only the underlying parameters are not known. For example, suppose we observe the dynamics of a simple dissipative linear spring, then we may express the model as $m\ddot{x} + \gamma\dot{x} + kx = 0$ based on Hooke's law. However, the parameters m, γ, and k might be unknown and need to be estimated in order to completely specify the model for purposes such as prediction and control. One may directly measure those parameters by static testing (e.g., weighing the mass on a scale). Alternatively, here we are interested in utilizing the observational data generated by the system without having to design and perform additional experiments to estimate the parameters corresponding to the model that best fits empirical observations, which is a standard viewpoint in SID. In this thought experiment, the SID process is performed with the underlying physics understood (the form of the Hooke spring equation). In general, it can be applied in the scenario where very little information was previously known about the system, in a black box manner.

Suppose that observations $\{\boldsymbol{z}(t)\}$ come from a general (multidimensional, coupled) DE, represented by

$$\dot{\boldsymbol{z}} = \boldsymbol{F}(\boldsymbol{z}), \tag{1}$$

where $\boldsymbol{z} = [z_1, \dots, z_N]^T \in \mathbb{R}^N$ is the (multivariate) state variable of the system and $\boldsymbol{F} = [F_1, \dots, F_N]^\top : \mathbb{R}^N \to \mathbb{R}^N$ is the vector field. Each component function $F_i(\boldsymbol{z})$ can be represented using a series expansion (for example, a power series or a Fourier series), written generally,

$$\dot{z}_i = F_i(\boldsymbol{z}) = \sum_{k=0}^{\infty} a_{ik}\phi_k(\boldsymbol{z}), \tag{2}$$

for a linear combination of basis functions $\{\phi_k\}_{k=0}^{\infty}$. The basis functions do not need to be mutually orthogonal, and the series can even include multiple bases, for example, to contain both a polynomial basis and a Fourier basis.[2] The coefficients $\{a_{ik}\}$ are to be determined by contrasting simulations to experimental measurements, in an optimization process whose details of how error is measured distinguish the various methods we discuss here. This was the main theme in previous approaches on nonlinear SID, with different methods differing mainly on how a model's fit is quantified. The different approaches include using standard squared error measures,[3,4] sparsity-promoting methods,[2,5–7] as well as using entropy-based cost functions.[8] Among those, sparsity-promoting methods have proven particularly useful because they tend to avoid the issue of overfitting, thus allowing a large number of basis functions to be included to capture possibly rich dynamical behavior.[2,5,6]

Regardless of the particular method or system, most previous work on nonlinear SID focused on the low-noise regime and demonstrated success only when there is a sufficient amount of clean observational data. In practice, an observation process can be subject to external disturbances in unpredictable ways. Consequently, the effective noise can be quite large and even with frequently occurring "outliers" both of which may contaminate the otherwise perfect data. Can SID still work under the presence of large noise and outliers? At a glance, the answer should be yes, given that several recent SID methods for nonlinear systems are readily deployable in the presence of noise. For example, compressive sensing can handle noise by relaxing the constraint set, whereas least squares and Lasso can be applied off the shelf—the important question, however, is whether the quality of solution is compromised or not, and to what extent. Recently, Tran and Ward considered the nonlinear SID problem under the presence of outliers in observational data and showed that so long as the outliers are "sparse" leaving sufficient amount of "clean" data available, existing techniques such as Sparse Identification of Nonlinear Dynamics (SINDy) can be extended to reconstruct the exact form of a system with high probability.[9] In the current work, we are interested in the more realistic scenario where effective noise is present everywhere and thus *all* data points are contaminated by non-negligible noise and sometimes outliers. These features effectively create a "high noise and low data amount" regime, where we found that existing nonlinear SID methods including recent ones that specialize in promoting sparsity fall short.

In this work, we depart from most standard approaches for nonlinear SID. We identify the error quantification via metric-based cost functions as a root cause of existing methods to fail under large noise and outliers because outliers tend to deviate from the rest of sample data as measured by metric distance; thus trying to "fit" the outliers almost inevitably causes the model to put (much) less weights on the "good" data points. To resolve this important issue, we propose to infer the (sparsity) structure of a general model together with its parameters using a novel *information-theoretic regression* approach that we call Entropic Regression (ER). As we will show, while standard metric-based methods emphasize the data in ways as designed by the chosen metric, the proposed ER approach is robust with regard to the presence of noise and outliers in the data. Instead of searching for the sparsest model and thus risk forcing a wrong sparse model, ER is emphasizing "information relevance" according to a model-free, entropic criterion. Basis terms will be included in the model only because they are relevant and not (necessarily) because they together make up the sparsest model. We demonstrate the effectiveness of ER in several examples, including chaotic Lorenz systems, Kuramoto-Sivashinsky (KS) equations, and a double-well potential, where in each case, the observed data contain relatively large noise and outliers. We also remark on the computational complexity and convergence in small-data regime, as well as discuss open problems and future directions.

Published under license by AIP Publishing.

RESULTS

Nonlinear system identification: Problem statement and formulation

Following the standard routine in nonlinear SID,[10] the starting point is to recast the nonlinear SID problem into a computational inverse problem, by considering an appropriate set of basis functions that span the space of functions including the system of interest.[3,7] A common choice is the standard *polynomial basis*

$$\begin{aligned}\boldsymbol{\phi} &= [\phi_0(\boldsymbol{z}), \phi_1(\boldsymbol{z}), \phi_2(\boldsymbol{z}), \dots] \\ &= [1, z_1, z_2, \dots, z_N, z_1z_2, z_1z_3, \dots, z_{N-1}z_N, \dots],\end{aligned} \tag{3}$$

where each term is a monomial. Using a set of basis functions, one can represent the individual component functions of F as a series as in (2). The specification of the location of nonzero parameters is referred to as the *structure* of the model.

Consider time series data $\{\boldsymbol{z}(t) = [z_1(t), \dots, z_m(t)]^\top\}_{t=t_0,\dots,t_\ell}$ and corresponding $\{\boldsymbol{F}(\boldsymbol{z}(t)\}_{t=t_0,\dots,t_\ell}$ generated from a nonlinear, high-dimensional dynamical system (1), possibly subject to observational noise. From $\boldsymbol{z}(t)$, one can estimate the derivatives by any of the standard Newton-Cotes methods, explicit Euler's method of course being the simplest, giving $F_i(\boldsymbol{z}(t_k)) = \frac{z_i(t_{k+1})-z_i(t_k)}{\tau_k} + \mathcal{O}((t_{k+1}-t_k))$ or central difference which has improved accuracy: $F_i(\boldsymbol{z}(t_k)) = \frac{z_i(t_{k+1})-z_i(t_{k-1})}{t_{k+1}-t_{k-1}} + \mathcal{O}((t_{k+1}-t_{k-1})^2)$. The problem of nonlinear system identification is to reconstruct the functional form as well as parameters of the underlying system, that is, to infer the nonlinear function $\boldsymbol{F}$.

Under the basis representation (2), the identification of $\boldsymbol{F}$ becomes equivalent to estimating all the parameters $\{a_{ik}\}$. In practice, the empirically observed state variable is subject to noise: $\hat{\boldsymbol{z}}(t) = \boldsymbol{z}(t) + \boldsymbol{\eta}(t)$ with $\boldsymbol{\eta}(t)$ representing the (multivariate) noise and $\hat{F}_i$ denoting the approximated value of F_i. For noisy observations $\hat{\boldsymbol{z}}(t)$, the difference between $\hat{F}_i(\hat{\boldsymbol{z}}(t))$ and $F_i(\hat{\boldsymbol{z}}(t))$ originates from several sources: the infinite series is truncated and the derivatives are estimated numerically and by using approximate states. Nevertheless, we can represent the aggregated error as an effective noise $\boldsymbol{\xi}(t)$ term and express the forward model as

$$\hat{F}_i(\hat{\boldsymbol{z}}(t)) = \sum_{k=0}^{K} a_{ik}\phi_k(\hat{\boldsymbol{z}}(t)) + \xi_i(t) \quad (t = t_0, \dots, t_\ell; i = 1, \dots, N). \tag{4}$$

Note that because of the combined and accumulated effects of observational noise, approximation error, and truncation, even if the observational noise of the states $\eta_i(t)$ is iid, this is not necessarily true for the effective noise $\xi_i(t)$. In matrix form, forward model (4) has the approximate expression

$$\begin{pmatrix} | & & | \\ \dot{z}_1(t_i) & \dots & \dot{z}_N(t_i) \\ | & & | \end{pmatrix} \approx \begin{pmatrix} | & | & & | \\ \phi_0(t_i) & \phi_1(t_i) & \dots & \phi_K(t_i) \\ | & | & & | \end{pmatrix} \times \begin{pmatrix} a_{00} & a_{01} & \dots & a_{0N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{K0} & a_{K1} & \dots & a_{KN} \end{pmatrix}. \tag{5}$$

Figure 1 shows the structure of the Lorenz system under standard polynomial basis up to quadratic terms.

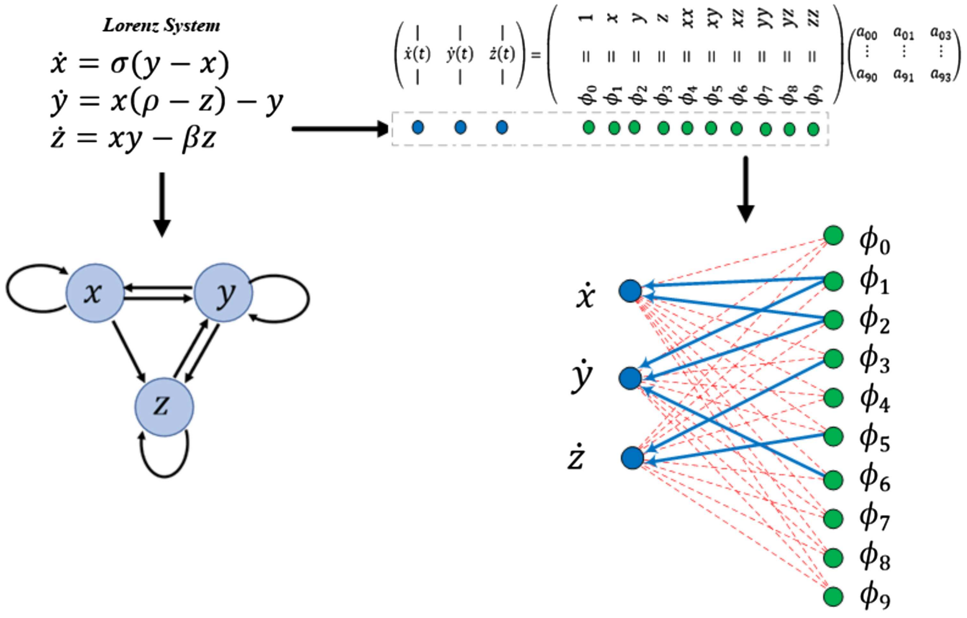

FIG. 1. (Left) Lorenz system as a dynamical system and its standard graph representation. (Right) Linear combination of nonlinear basis functions, with coupling coefficients $\{a_{ik}\}$ forming the structure of the system (bottom right). Here, each directed edges represent the presence of basis terms on the individual variables of the system.

Published under license by AIP Publishing.

In vector form, under a choice of basis and truncation, the nonlinear system identification problem can be recast into the form of a linear inverse problem,

$$\boldsymbol{f}^{(i)} = \Phi \boldsymbol{a}^{(i)} + \boldsymbol{\xi}^{(i)}, \tag{6}$$

where $\boldsymbol{f}^{(i)} = [\hat{F}_i(\hat{\boldsymbol{z}}(t_1)), \dots, \hat{F}_i(\hat{\boldsymbol{z}}(t_\ell))]^\top \in \mathbb{R}^{\ell\times 1}$ represents the ith component of the estimated vector field from the observational data, $\Phi = [\boldsymbol{\phi}^{(1)}, \dots, \boldsymbol{\phi}^{(K)}] \in \mathbb{R}^{\ell\times K}$ [with $\boldsymbol{\phi}^{(k)} = [\phi_k(\hat{\boldsymbol{z}}(t_1)), \dots, \phi_k(\hat{\boldsymbol{z}}(t_\ell))]$ $\in \mathbb{R}^{\ell\times 1}$] represent sampled data for the basis functions, $\boldsymbol{\xi}^{(i)} = [\xi_i(t_1), \dots, \xi_i(t_\ell)]^\top \in \mathbb{R}^{\ell\times 1}$ represents noise, and $\boldsymbol{a}^{(i)}$ $= [a_{i1}, \dots, a_{iK}]^\top \in \mathbb{R}^{K\times 1}$ is the vector of parameters, which is to be determined. Note that the form of Eq. (6) is the same for each i, and solving each $\boldsymbol{a}^{(i)}$ can be done separately and independently for each i. In what follows, we omit the index when discussing the general methodology and consider the following linear inverse problem:

$$\boldsymbol{f} = \Phi \boldsymbol{a} + \boldsymbol{\xi}, \tag{7}$$

where $\boldsymbol{f} \in \mathbb{R}^{\ell\times 1}$ and $\Phi \in \mathbb{R}^{\ell\times K}$ are given, with the goal to estimate $\boldsymbol{a} \in \mathbb{R}^{K\times 1}$. This general problem is in the form of an inverse problem and is typically solved under various assumptions of noise by methods such as least squares, orthogonal least squares (OLS), lasso, and compressed sensing, to name a few. Each of these methods, in addition to the recent approach of SINDy and its generalization, is mentioned in the Results section and reviewed in the Methods section. In what follows we develop a unique information-theoretic approach called entropic regression, which we demonstrate has significant advantages.

Entropic regression

To overcome the competing challenges of potential overfitting, efficiency when limited data points are given, and robustness with respect to noise and, in particular, outliers in observations, we propose a novel framework that combines the advantage of information-theoretic measures and iterative regression methods. The framework, which we term *entropic regression* (ER), is model-free, noise-resilient, and efficient in discovering a "minimally sufficient" model to represent data. The key idea is that, for given set of basis functions, a model should be considered minimally sufficient if no basis function that is not already included in the model can help increase the information relevance between the model outputs and observed data. In other words, the residual between the model fit and observational data is statistically independent from any basis function that is not included in the model—because otherwise the dependence can be harvested to reduce the discrepancy by including such a basis function in the model. We emphasize that, although the idea seems related to classical model selection principles such as Akaike information criterion (AIC),[11] ours combines model construction with selection. In addition, even though it is not uncommon for entropy measures to be adopted in system identification,[8,12] the proposed method is unique as it fuses entropy optimization with regression in a principled manner that enables scalable computation and efficient estimation in reconstruction nonlinear dynamics under noisy data. As we shall see below, the proposed ER method is applicable even in the small-sampling regime (by adopting appropriately defined entropy measures and efficient estimators) and naturally allows for a computationally efficient procedure to build up a model from scratch. In particular, we use (conditional) mutual information as an information-theoretic criterion and iteratively select relevant basis functions, analogous to the optimal causation entropy algorithm previously developed for causal network inference[13,14] but here including an additional regression component in each step. Thus, ER can be thought of as an information-theoretic extension of the orthogonal least squares regression or as a regression version of optimal causation entropy.

We now present the details of ER. The ER method contains two stages (also see Algorithm 1 for the pseudocode): forward ER and backward ER. In both stages, selection and elimination are based on an entropy criterion and parameters are updated in each iteration using a standard regression (e.g., least squares). Consider the inverse problem (7). For an index set $S \subset \mathbb{N} \cup \{0\}$, the estimated parameters can be thought of as a mapping from the joint space of $\Phi, \boldsymbol{f}$, and S to a vector denoted as $\hat{\boldsymbol{a}} = R(\Phi, \boldsymbol{f}, S)$. For instance, under a least squares criterion, the mapping is given by $R(\Phi, \boldsymbol{f}, S)_S = \Phi_S^\dagger \boldsymbol{f}$ (Φ_S denotes the columns of matrix Φ indexed by S) and $R(\Phi, \boldsymbol{f}, S)_i = 0$ for all $i \notin S$. Using the estimated parameters, the recovered signal can be computed as $\Phi R(\Phi, \boldsymbol{f}, S)$. In the ER algorithm, we start by selecting a basis function $\boldsymbol{\phi}_{k_1}$ that maximizes its mutual information with $\boldsymbol{f}$, compute the corresponding parameter a_{k_1} using the least squares method, and obtain the corresponding regression model output $\boldsymbol{z}_1$ according to

$$\begin{cases} k_1 = \arg\max_k I(\Phi R(\Phi, \boldsymbol{f}, \{k\}); \boldsymbol{f}), \\ \hat{\boldsymbol{a}} = R(\Phi, \boldsymbol{f}, k_1), \\ \boldsymbol{z}_1 = \Phi R(\Phi, \boldsymbol{f}, k_1). \end{cases} \tag{8}$$

Here, $I(\boldsymbol{x}; \boldsymbol{y})$ denotes mutual information between $\boldsymbol{x}$ and $\boldsymbol{y}$, which is a model-free measure of the statistical dependence between two distributions (that is, $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent if and only if their mutual information equals zero; however, in practice, due to finite samples and statistical estimation, we wish to distinguish that Mutual Information (MI) is statistically insignificantly indistinguishable from zero and noting that it is never negative as well).[15] Next, in each iteration of the forward stage, we perform the following computations and updates for $i = 2, 3, \dots$:

$$\begin{cases} k_i = \arg\max_{k \notin \{k_1, \dots, k_{i-1}\}} I(\Phi R(\Phi, \boldsymbol{f}, \{k\}); \boldsymbol{f} | \boldsymbol{z}_{i-1}), \\ \hat{\boldsymbol{a}} = R(\Phi, \boldsymbol{f}, \{k_1, \dots, k_i\}), \\ \boldsymbol{z}_i = \Phi R(\Phi, \boldsymbol{f}, \{k_1, \dots, k_i\}). \end{cases} \tag{9}$$

The process terminates when $\max_k I(\Phi R(\Phi, \boldsymbol{f}, k); \boldsymbol{f} | \boldsymbol{z}_{i-1}) \approx 0$ (or when all basis functions are exhausted), indicating that none of the remaining basis function is *relevant* given the current model, in an information-theoretic sense. The result of the forward ER is a set of indices $S = \{k_1, \dots, k_m\}$ together with the corresponding parameters $a_{k_1}, \dots, a_{k_m}$ ($a_j = 0$ for $j \notin S$) and model $f \approx a_{k_1}\boldsymbol{\phi}_{k_1} + \cdots + a_{k_i}\boldsymbol{\phi}_{k_i}$. Finally, we turn to the backward stage, where the terms that had previously been included are re-examined for their information-theoretic relevance and those that are redundant will be removed. In particular, we sequentially check for each $j = k_i \in S$ to determine if the basis term $\boldsymbol{\phi}_j$ is redundant by computing

$$\begin{cases} \hat{\boldsymbol{a}} = R(\Phi, \boldsymbol{f}, \{k_1, \dots, k_i\}/\{k_i\}), \\ \bar{\boldsymbol{z}}_j = \Phi R(\Phi, \boldsymbol{f}, \{k_1, \dots, k_i\}/\{k_i\}), \end{cases} \tag{10}$$

Published under license by AIP Publishing.

and updating $S \to S/\{j\}$ (that is, remove j from the set S) if $I(\Phi R(\Phi, \boldsymbol{f}, S); \boldsymbol{f}|\bar{z}_j) \approx 0$. The result of the backward ER is the reduced set of indices $S = \{\ell_1, \dots, \ell_n\}$ with $n \leq m$, together with the corresponding parameters $a_{\ell_1}, \dots, a_{\ell_n}$ ($a_j = 0$ for $j \notin S$) computed as $\boldsymbol{a} = R(\Phi, \boldsymbol{f}, S)$, and accordingly the recovered model $\boldsymbol{f} \approx \boldsymbol{\phi}$ $\boldsymbol{a} = \boldsymbol{\phi}_S \boldsymbol{a}_S = a_{\ell_1}\boldsymbol{\phi}_{\ell_1} + \cdots + a_{\ell_n}\boldsymbol{\phi}_{\ell_n}$. In practice, mutual information and conditional mutual information need to be estimated from data, and whether or not the estimated values should be regarded as zero is typically done via (approximate) significance testing, the details of which are provided in *Methods* section (also see Supplementary Materials).

Algorithm 1. Entropic regression

```
procedure Initialization: $(\boldsymbol{f}, \Phi)$
        Tolerance ($tol$) Estimation.
        For a set of index $S$, define the function $R(\Phi, \boldsymbol{f}, S) = \Phi_S^{\dagger}\boldsymbol{f}$
  end procedure
  procedure Forward ER: $(\boldsymbol{f}, \Phi, tol)$
          $S_f = \emptyset, p = \emptyset, v = \infty, z = \emptyset$
          While $v > tol$ do
          $S_f \leftarrow p$
            $I_j^{est} := I(\Phi R(\Phi, \boldsymbol{f}, \{S_f, j\}); \boldsymbol{f}|z)$. for all $j \notin S_f$
              $v, p := \max_j I_j^{est}$
            $\hat{\boldsymbol{a}} := R(\Phi, \boldsymbol{f}, \{S_f, p\}))$
            $z := \Phi\hat{\boldsymbol{a}}$
            end while
            return $S_f$
   end procedure
   procedure BACKWARD ER:$(\boldsymbol{f}, \Phi, tol, S_f)$
            $S_b = S_f, p = \emptyset, v = -\infty$
            while $v < tol$ do
              $S_b := \{S_b\} - \{p\}$
              for all $j \in S_b$ do
                $\hat{\boldsymbol{a}} := R(\Phi, \boldsymbol{f}, \{S_b\} - \{j\}))$
                $z := \Phi\hat{\boldsymbol{a}}$
                $I_j^{est} := I(\Phi R(\Phi, \boldsymbol{f}, S_b); \boldsymbol{f}|z)$,
              end for
              $v, p := \min_j(I_j^{est})$
            end while
            return $S_b$
   end procedure
   return $S = S_b$.
```

Numerical experiments: Outliers, expansion order, and the paradox of sparsity

To demonstrate the utility of ER for nonlinear system identification under noisy observations, we benchmark its performance against existing methods including least squares (LS), orthogonal least squares (OLS), Lasso, as well as SINDy and its extension by Tran and Ward (TW). The details of the existing approaches are described in the Methods section. The examples we consider represent different types of systems and scenarios, including both Ordinary Differential Equations (ODEs) and Partial Differential Equations (PDEs). In addition, we consider different noise models and especially the presence of outliers in order to evaluate the robustness of the respective methods.

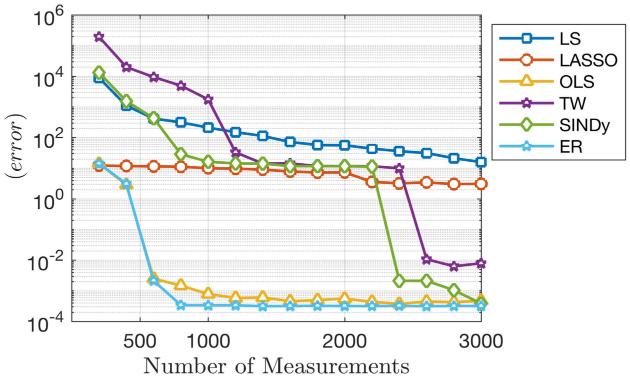

FIG. 2. Lorenz system. We perform 100 runs for the comparison, no outliers, 0.0005 step size, and we considered the median result out of 100 runs. The figure shows the error in the parameter estimation for a Lorenz system but subject to noisy measurements by Gaussian noise, with $\varepsilon = 10^{-4}$, and using a 5th-order polynomial expansion. We see that ER and OLS have an overall superior performance compared to other standard methods. We see that SINDy and TW are less successful (under a large span of tuning parameters, see Fig. 3) at this number of measurements even with low noise levels.

For each example system, we sample the state of each variable at a uniform rate of Δt to obtain a multivariate time series $\{\boldsymbol{z}(t_i)\}_{k=1,\dots,N; i=1,\dots,\ell}$, where $\boldsymbol{z} = [z_1, \dots, z_d]^{\top} \in \mathbb{R}^d$; then, we add noise to each state variable and obtain the noisy empirical time series

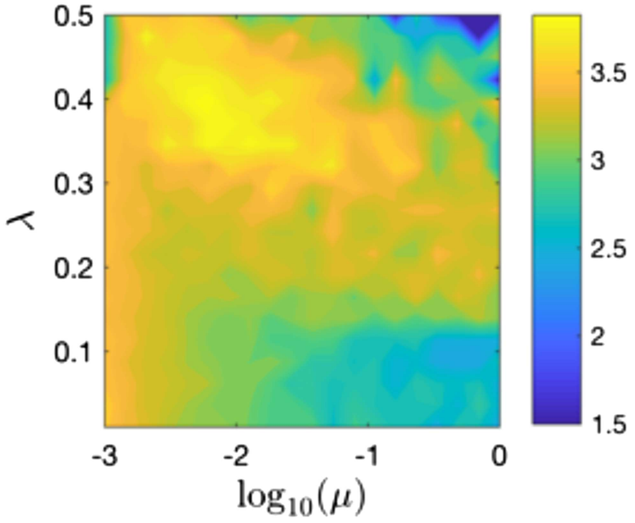

FIG. 3. Contour plot of the error in recovered solution of the Lorenz system (Fig. 2) by the TW method for a grid of μ and λ values and using 2000 measurements, 5th order polynomial expansion, low noise with $\varepsilon_1 = 10^{-5}$, and no corrupted data. The color bar indicates the value of $\log_{10}(error)$ in the recovered solution, and it shows large error at all levels of tuning parameters.

Published under license by AIP Publishing.

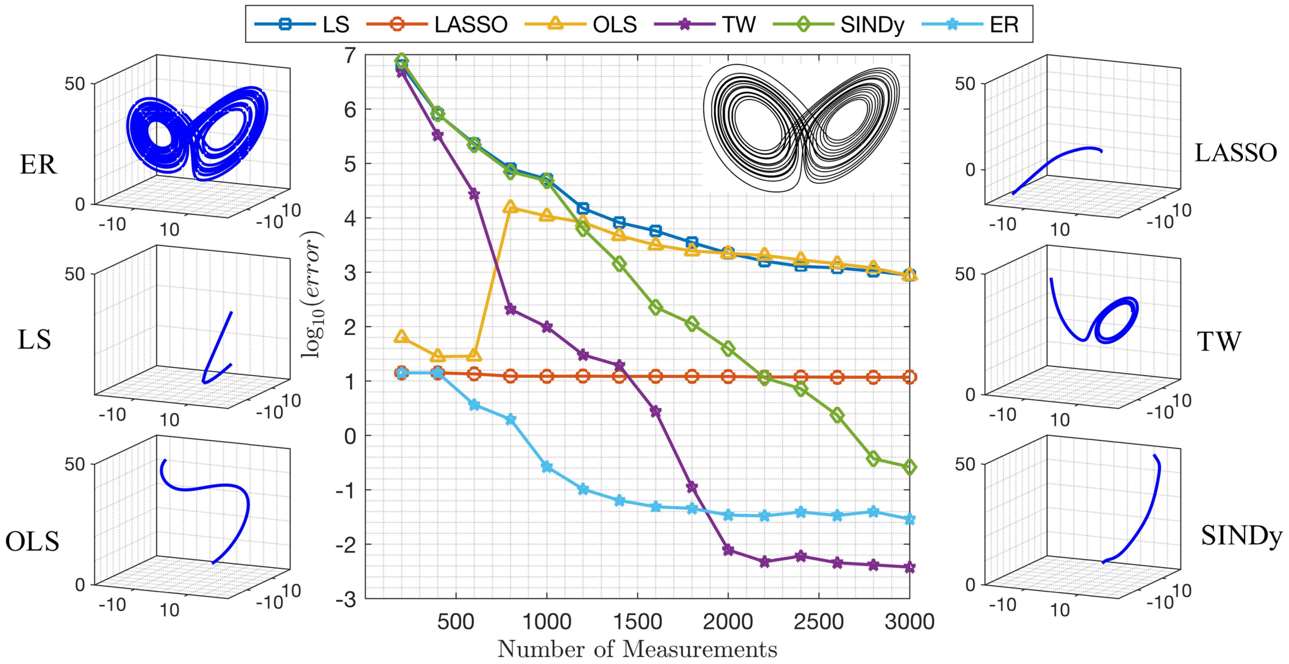

FIG. 4. SID for the Lorenz system when the observations are corrupted by outliers. Contrast to Fig. 2. As before, we specify a level of persistent Gaussian observation noise, $\eta \sim \mathcal{N}(0, \varepsilon_1)(1 - Ber(p))$, but now furthermore we allow for an "outlier noise," as "occasional" bursts of much larger perturbations, $\eta \sim \mathcal{N}(0, \varepsilon_1 + \varepsilon_2) Ber(p)$, where $Ber(p)$ is the standard Bernoulli random variable (0 or 1 with probability ratio p, and $0 \leq p \leq 1$). (Middle) Error in estimated parameters for the Lorenz system given in Eq. (12) with noise, $\varepsilon_1 = 10^{-5}$, $\varepsilon_2 = 0.2$, 5th-order polynomial expansion, and $p = 0.2$. The Lorenz system dynamics is shown in the upper right corner. We see that ER has fast convergence at a low number of measurements, followed by TW which required twice the number of measurements. Different from TW, in our ER method, we focus on detecting the true sparse structure with the presence of outliers, without any attempts to neglect outliers based on some weight function to achieve higher accuracy, which is the case in the TW method. This point clearly appears in Fig. 5 where we see that although TW achieved higher accuracy, it has a low exact recovery probability, while ER reached exact recovery probability more than 90%. A detailed statistics over the 100 runs is discussed in the supplementary material. (Side panels) Typical trajectories generated by the reconstructed dynamical systems, where for each method, we show results using the "median" solution, that is, the recovered system whose corresponding parameter estimation error is at the median over a large number of independent simulation runs. In each such simulation, 1500 samples are used. Comparing with the true Lorenz attractor (upper right corner in the main panel), we see that the only reasonable reconstruction in this case was produced by ER.

denoted by $\{\hat{\boldsymbol{z}}(t_i)\}$, where

$$\hat{z}_k(t_i) = z_k(t_i) + \eta_{ki}, \tag{11}$$

with η_{ki} representing state observational noise. The vector field $\boldsymbol{F}$ is estimated using the central difference on the noisy time series $\{\hat{\boldsymbol{z}}(t)\}$.

Example 1. *Chaotic Lorenz system.* Our first detailed example data set was generated by noisy observations from a chaotic Lorenz system, which is represented by a three-dimensional ODE that is a prototype system as a minimal model for thermal convection obtained by a low-ordered modal truncation of the Saltzman PDE[16] and for many parameter combinations exhibits chaotic behavior.[17] In our standard notation, we have $\boldsymbol{z} = [z_1, z_2, z_3]^\top$ and

$$\begin{cases} \dot{z}_1 = F_1(\boldsymbol{z}) = \sigma(z_2 - z_1), \\ \dot{z}_2 = F_2(\boldsymbol{z}) = z_1(\rho - z_3) - z_2, \\ \dot{z}_3 = F_3(\boldsymbol{z}) = z_1 z_2 - \beta z_3, \end{cases}$$

with default parameter values $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$ unless otherwise specified. We consider a standard polynomial basis as in Eq. (3). Over recent years, the Lorenz system has become a favorable and standard example for testing SID methods and typically requires tens of thousands of measurements for accurate reconstruction.[2,9]

First, we compare several nonlinear SID methods in reconstructing the Lorenz system when the state observational noise is drawn independently from a Gaussian distribution, $\eta \sim \mathcal{N}(0, \varepsilon^2)$. As we discussed before, this translates into effective noise that is not necessarily Gaussian or even independent. Figure 2 shows the error in the estimated parameters, where $error = \|\boldsymbol{a}_{true} - \boldsymbol{a}_{estimated}\|_2$. As shown in Fig. 2, even with observational noise as low as $\varepsilon = 10^{-4}$, ER and OLS outperform all other methods. In this low-noise regime, SINDy required more measurements (around 4 times) to reach similar accuracy as ER. In comparison, as noted in Refs. 9 and 2 and in the implementation provided by the authors, for SINDy and TW methods to yield accurate reconstruction, the number of measurements is at the order of 10^4. See Fig. 3 for the results of the TW method with a large span of tuning parameters.

Next, to explore the performance of SID methods under the presence of outliers, we conduct additional numerical experiments. The extent to which outliers present is controlled by a single parameter p: each observation is subject to an added noise η, where

Published under license by AIP Publishing.

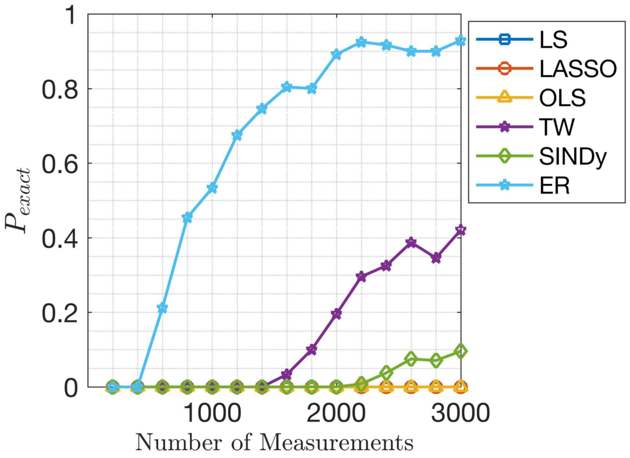

FIG. 5. Probability of exact recovery for the Lorenz system. For the same results shown in Fig. 4, *Pexact* here represents the number of runs in which a method recovered the exact sparse structure over the total number of runs. We see that although TW reached high accuracy at a high number of measurements, its exact recovery probability remains low.

$\eta \sim \mathcal{N}(0, \varepsilon_1^2)$ with probability $1 - p$ and $\eta \sim \mathcal{N}(0, \varepsilon_1^2 + \varepsilon_2^2)$ with probability p. Here, we use $\varepsilon_1 = 10^{-5}$, $\varepsilon_2 = 0.2$, and $p = 0.2$. The results of SID are shown in Figs. 4 and 5. Compared to Fig. 2, we see that with $p > 0$, OLS performance drops due to the increasing occurrence of large noise and outliers, whereas ER retains its capacity of accurately identifying the underlying system. As an example, in each of the side panels of Fig. 4, we show the trajectory of the identified dynamics using the median solution of each method. It is clear that under such noisy chaotic dynamics and at a relatively undersampled regime, the ER method successfully recovers the system dynamic. As an ample amount of data becomes available, we note that the TW method starts to produce excellent reconstruction, which is consistent with recent findings reported in Ref. 9.

Given that a major theme of modern SID is to seek for *sparse* representations and the Lorenz system under standard polynomial basis is indeed sparse, it is worth asking: what are the respective structures identified by the different methods? In Fig. 6, we compare the structure of the identified model using different methods across a range of parameter values for ρ. In this case, under the presence of large noise and outliers ($p = 0.2$), none of the methods examined here, including recently proposed sparsity-promoting (CS, SINDy) and outlier-resilient (TW) methods, is able to identify the correct structure. The proposed ER method, however, does identify the correct structure. It is worth pointing out that, often times when expressed in the right basis, a model will appear to be sparse, the converse is not true: just because a method return a sparse solution does not suggest (at all) the such a solution gives a reasonable approximation of the true model structure. Interestingly, as we discuss in the supplementary material, for the same system and data, as more basis functions are used—that is, when the true dynamics becomes sparser—the

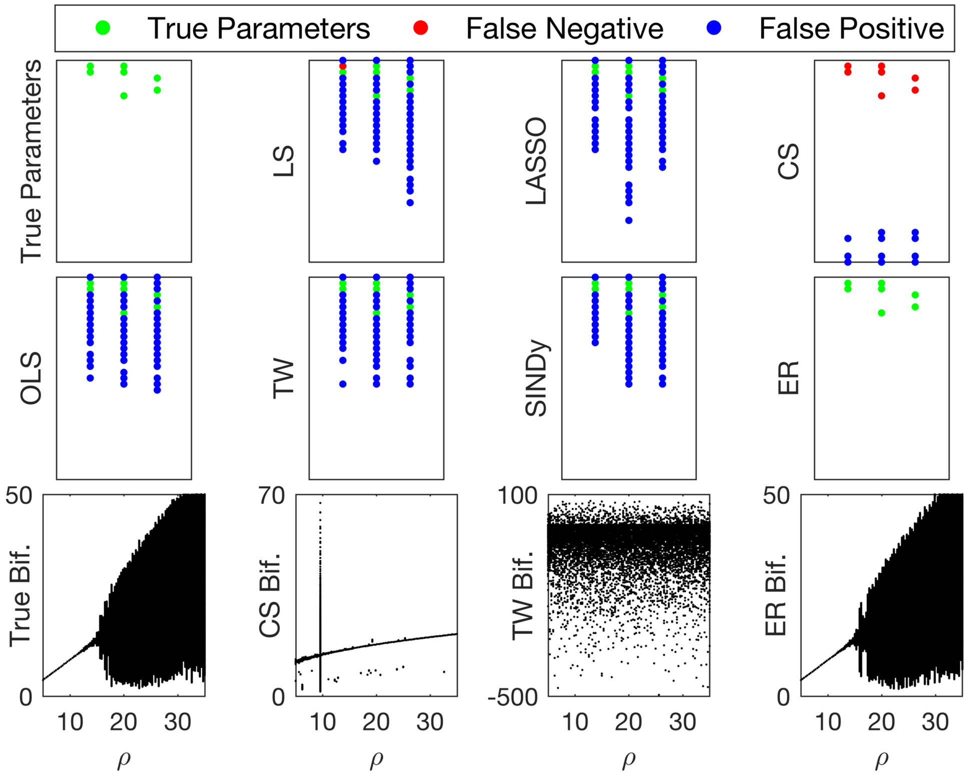

FIG. 6. Sparse representation of the solution found by solvers using 1500 measurements, and $p = 0.2$ on Fig. 4. The upper left corner shows the true solution of the Lorenz system. The bottom column shows the bifurcation diagram on z dimension of the Lorenz system with $\rho \in [5, 30]$ as the bifurcation parameter, created using 5000 initial conditions evolved according the recovered solution.

Published under license by AIP Publishing.

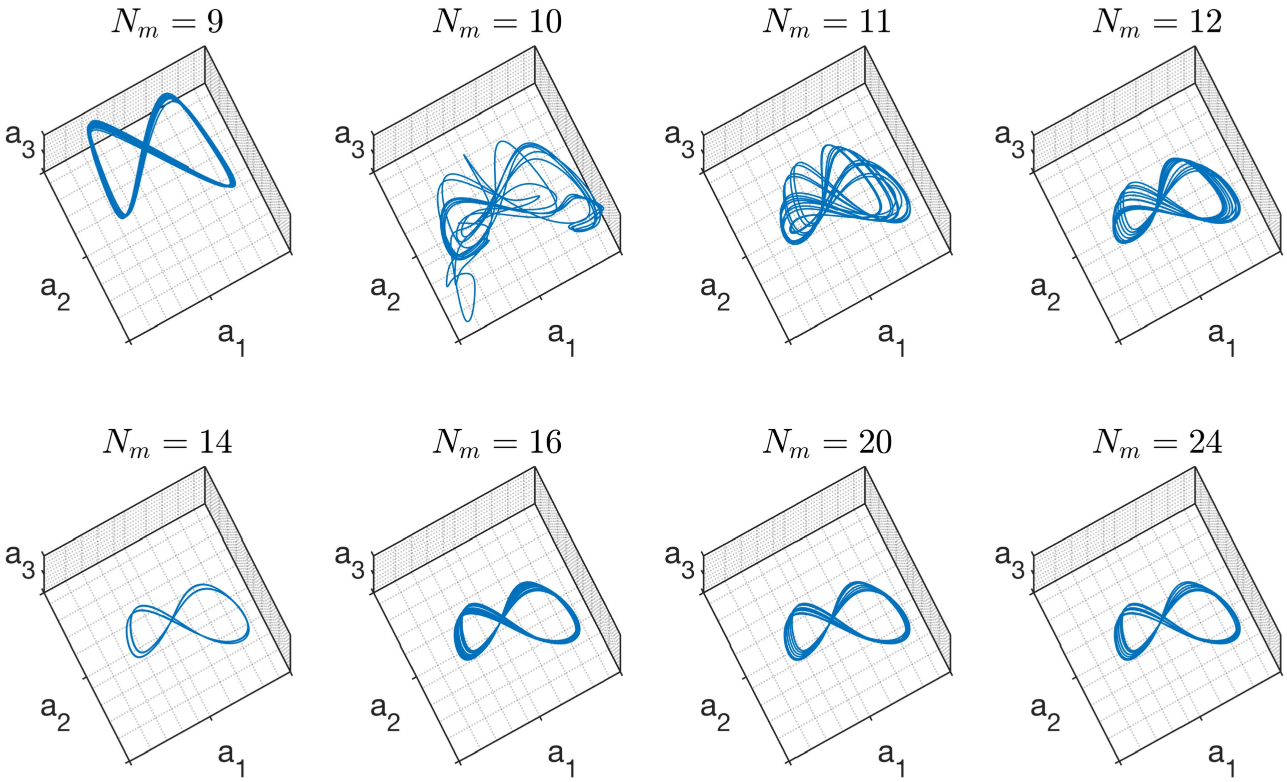

FIG. 7. The first three modes of the ODE equation (15) solution. We show the modes a_1, a_2, and a_3 for the selected number of modes. For a clear view, we fixed the axis limits to be $a_1 \in [-1.21, 1.06]$, $a_2 \in [-0.75, 0.98]$, and $a_3 \in [-1.1, 1.12]$ for all plots. We found that there was no significant addition to the dynamic with $16 < N_m$ (meaning that $N_m = 16$ was enough to describe the system).

reconstructed dynamics using existing methods (such as CS) can become worse.

Example 2. (**Kuramoto-Sivashinsky equations**). To further demonstrate the power of ER, we consider a nonlinear PDE, namely, the Kuramoto-Sivashinsky (KS) equation,[18–22] which arises as a description of flame front flutter of gas burning in a cylindrically symmetric burner. It has become a popular example of a PDE that exhibits chaotic behavior, in particular, spatiotemporal chaos.[23,24] We will consider the Kuramoto-Sivashinsky system in the following form:

$$u_t = -\nu u_{xxxx} - u_{xx} + 2uu_x, \quad (t,x) \in [0,\infty) \times (0,L) \tag{12}$$

in periodic domain, $u(t,x) = u(t,x+L)$, and we restrict our solution to the subspace of odd solutions $u(t,-x) = -u(t,x)$. The viscosity parameter ν controls the suppression of solutions with fast spatial variations and is set to $\nu = 0.029910$ under which the system exhibits chaotic behavior.[23]

Since a PDE corresponds to an infinite-dimensional dynamical system, in practice, we focus on an approximate finite-dimensional representation of the system, for example, by Galerkin projection onto basis functions as infinitely many ODEs in the corresponding Banach space.

To develop the Galerkin projection, we follow the procedure as presented in Ref. 25, to expand a periodic solution $u(x,t)$ using a discrete spatial Fourier series,

$$u(x,t) = \sum_{-\infty}^{\infty} b_k(t) e^{ikqx}, \tag{13}$$

where $q = \frac{2\pi}{L}$.

Notice that we have written this Fourier series of basis elements e^{ikqx} in terms of time varying combinations of basis elements. For simplicity, consider $L = 2\pi$, then $q = 1$ for the following analysis. This is typical[26] with the representation of a PDE as infinitely many ODEs in the Banach space, where orbits of these coefficients, therefore, become time varying patterns by Eq. (13). Substituting Eq. (13) into Eq. (12), we produce the infinitely many evolution equations for the Fourier coefficients,

$$\dot{b}_k = (k^2 - \nu k^4) b_k + ik \sum_{m=-\infty}^{\infty} b_m b_{k-m}. \tag{14}$$

In general, the coefficients b_k are complex functions of time t. However, by symmetry, we can reduce to a subspace by considering the special symmetry case that b_k is purely imaginary, $b_k = ia_k$ and

Published under license by AIP Publishing.

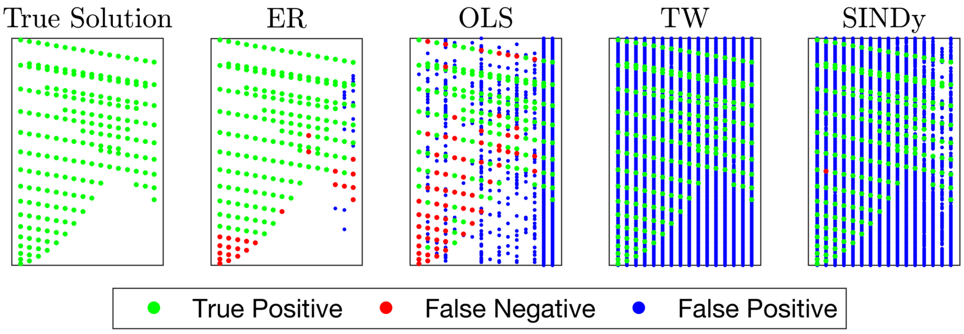

FIG. 8. In analogy to Fig. 6, sparse representation of the KSE solution by different methods. CS and LASSO have been excluded for their high computation complexity.

$a_k \in \mathbb{R}$. Then,

$$\dot{a}_k = (k^2 - \nu k^4)a_k - k \sum_{m=-\infty}^{\infty} a_m a_{k-m}, \tag{15}$$

where $k = 1, \ldots, N_m$. However, the assumption that there is a slow manifold (slow modes as an inertial manifold[26–29]) suggests the practical matter that a finite truncation of the series Eq. (13), and correspondingly the reduction to finitely many ODEs will suffice. Therefore, we choose a sufficiently large number of modes N_m. Then, we solve the resulting N_m-dimensional ODE (15) to produce the estimated solution of $u(x, t)$ by (13), and use such data for the purpose of SID, so as to estimate the structure and parameters of the ODE model (15).

Figure 7 shows the first three dimensions plot under different number of modes. We see that using just a few number of modes ($N_m = 8, \ldots, 11$) is insufficient to capture the true dynamical behavior of the system, whereas too large a number of modes ($N_m = 20, 24$) may be unnecessary. In this example, an adequate but not excessive number of modes seems to be around $N_m = 16$, as no significant information is gained by increasing N_m.

Figure 8 shows the sparse structure of the recovered solution by different methods. Here, we mention that the true nonzero parameters of Kuramoto-Sivashinsky equations (KSE) using $N_m = 16$ are 200 parameters that vary in the magnitude from 0.9701 to 1705. With the second order expansion, our basis matrix will have 153 candidate functions, and it will be nearly singular with condition number 4×10^7. Likely due to such high condition number, neither TW nor SINDy gives reasonable reconstruction. In particular, we note that the solution of SINDy is already optimized by selecting the threshold value λ that is slightly above λ_*, where $\lambda_* \approx 0.1731$ is the smallest magnitude of the true nonzero parameter of the full least squares solution. A larger value of λ only worsens the reconstruction, as we found numerically.

The OLS method overcomes the disadvantage of LS by iteratively finding the most relevant "feature" variables, where relevance is measured in terms of (squared) model error, but it comes at a price: similar to LS, the OLS is sensitive to outliers in the data and such sensitivity seems to be even more amplified due to the smaller number of terms typically included in OLS as compared to LS, which cause the high false negative rate in the OLS solution. Although the ER solution has a few false negatives, it was completely able to recover the overall dynamic of the system as shown in Fig. 9, while all other solutions diverges and failed to recover $u(x, t)$.

Example 3. (**Double-Well Potential**). Finally, in order to gain further insights into why standard methods fail under the presence of outliers, we consider a relatively simple double-well system, with

$$f(x) = x^4 - x^2. \tag{16}$$

Suppose that we measure x and f, can we identify the function $f(x)$? We sample 61 equally spaced measurements for $x \in [-1.2, 1.2]$, and we construct Φ using the 10th order polynomial expansion with

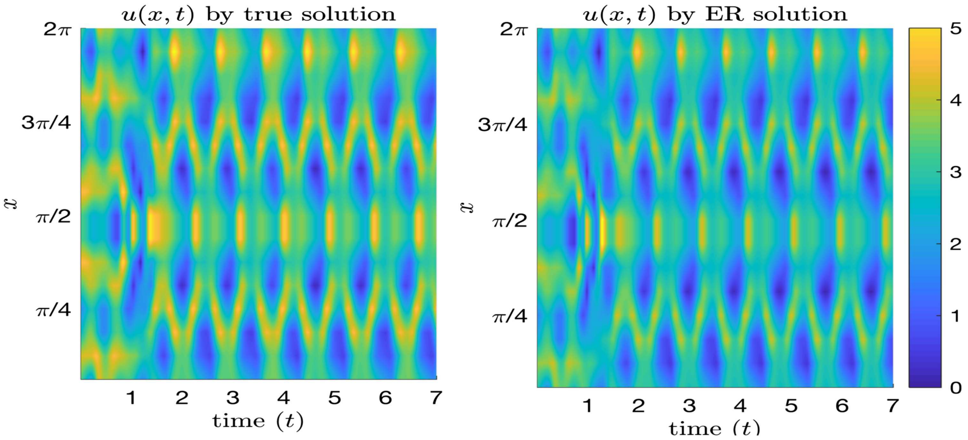

FIG. 9. $u(x, t)$ constructed by the true solution (left) and the ER solution (right) using Eq. (13). OLS and TW were not able to reproduce the dynamic and they diverge after a few iterations. We see that the reconstructed dynamic using ER solution is identical to the true solution with a minor difference in the transient time, although there was a false negative in the ER solution. ER detected the stiff parameters that dominate the overall dynamic. Sloppiness of some KSE parameters makes their influence practically negligible to the overall dynamic.

Published under license by AIP Publishing.

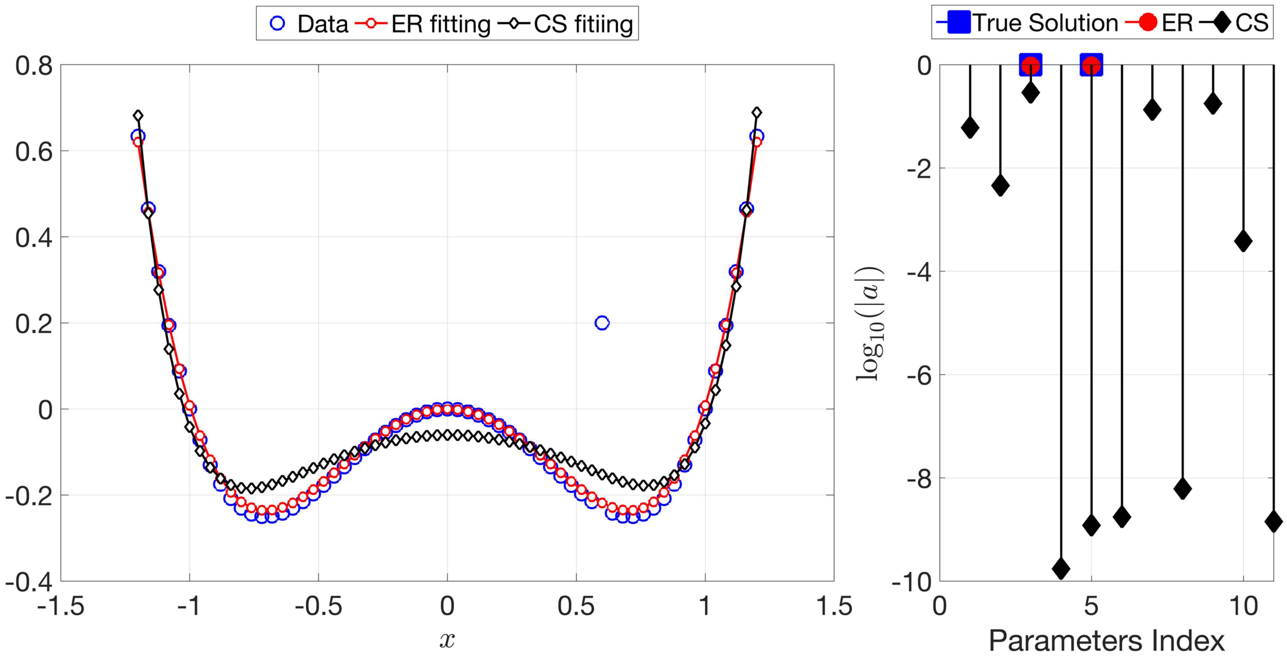

FIG. 10. Double-well potential given by Eq. (16) data fitting using ER and CS. CS solution found as the solution with minimum residual from 100 log-spaced values of $\varepsilon \in [10^{-9}, 10^{2}]$.

$K = 11$ being the number of candidate functions. Then, we consider a single fixed value corrupted measurement to be $f(0.6) = 0.2$.

Figure 10 shows the results the double-well SID under a single outlier in the observation. We see the robustness of ER solution to the outliers while CS failed in detecting the system sparse structure. For the sake of clearness, Fig. 10 shows the results for CS and ER. The results for each solver and details are provided in the supplementary material.

DISCUSSION

The main theme of the paper is on nonlinear system identification (SID) under noisy observations, which is to learn the functional form and parameters of a nonlinear system based on observations of its states under the presence of noise and outliers. We recast the problem into the form of an inverse problem using a basis expansion of the nonlinear functions. Such basis expansion, however, renders the resulting problem inherently high dimensional even for low-dimensional systems. In practice, the need for finite-order truncation as well as the presence of noise causes additional challenges. For instance, even under iid Gaussian observational noise for the state variables, the effective noise in the inverse problem is not necessarily so. As we demonstrate using several example systems, including the chaotic Lorenz system and the Kuramoto-Sivashinsky equations, existing SID methods are prone to noise and can be quite sensitive to the presence of outliers. We identify the root cause of such nonrobustness to the metric nature of the existing methods, as they quantify error based on metric distance, and thus a handful of data points that are "corrupted" by large noise can dominate the model fit. Each of the existing methods we considered has this property, which includes the least squares, compressive sensing, and Lasso. From a mathematical point of view, each method can be interpreted as a functional that maps input data to a model, through some optimization process. In a noisy setting, the output model should ideally change smoothly with respect to the input data, not just continuously. Our results suggest that these popular methods in fact do suffer from a sensitive dependence on outliers, as a few corrupted data can already produce very poor model estimates. Alarmingly, the now-popular CS method, which is based on sparse regression, can force to select a completely wrong sparse model under noisy input data, and this occurs even when there is just a single outlier. This is by no means contradicting previous findings of the success of CS in SID, as in such work, noise is typically very small, and here we are considering a perhaps more realistic scenario with larger noise.

To fill the vacancy of SID methods that can overcome outliers, we develop an information-theoretic regression technique, called entropic regression (ER), that combines entropy measures with an iterative optimization for nonlinear SID. We show that ER is robust to noise and outliers, in the otherwise very challenging circumstances of finding a model that explains data from dynamical stochastic processes. The key to ER's success is its ability to recover the correct and true sparsity structure of a nonlinear system under basis expansions, despite either relatively large noise or alternatively even relatively many even larger outliers. In this sense, ER is superior to any other method that we know of for such settings. Note that in the ER algorithm, least squares is used to estimate the parameters of those

Published under license by AIP Publishing.

basis functions that are deemed relevant where relevance is detected using an information-theoretic measure that is insensitive to noise and outliers. The choice of least squares in the regression step in ER is not necessarily an optimal choice and can be potentially replaced by more advanced methods (e.g., those developed in robust regression). In the current implementation of ER, we adopted least squares mainly due to its computational advantage over alternative methods. On a more fundamental level, ER's robustness against outliers may likely be attributed to an important principle in information theory called the asymptotic equipartition property (AEP).[15] The outcome of this principle is that sampled data can be partitioned into "typical" samples and "atypical" samples, with the rare atypical samples ending up influencing the estimated entropy relatively weakly. Since ER measures relevance by entropy instead of metric distance, a few outliers, no matter how far away they are from the rest of the data points, tend to have minimal impact on the model identification process. So, the general interpretation we make here is that outliers observations are likely atypical, but not part of the core of data that carry the major estimation of the entropy. This foundational concept of information theory is likely the major source of robustness of our ER method to system identification.

METHODS

Existing metric-based methods for system identification

Recall (from the main text) that we recast the nonlinear system identification problem here. Given a truncated basis representation of each component of the vector field $\boldsymbol{F}$, expressed as

$$F_i(\boldsymbol{z}) = \sum_{k=0}^{K} a_{ik}\phi_k(\boldsymbol{z}), \tag{17}$$

we consider sampled data $\hat{\boldsymbol{z}}$ and the estimated vector field $\hat{\boldsymbol{F}}$, from which the coefficients (parameters) $\{a_{ik}\}$ are to be determined. In general, we use subscript "t" to index the sampled data, and thus the tth sample satisfies the equation

$$\hat{F}_i(\hat{\boldsymbol{z}}(t)) = \sum_{k=0}^{K} a_{ik}\phi_k(\hat{\boldsymbol{z}}(t)) + \xi_i(t) \quad (t = 1,\dots,T;\ i = 1,\dots,n). \tag{18}$$

Here, $\xi_i(t)$ is the effective noise that represents the accumulative impact of truncation error, state observational noise, as well as approximation error in the estimation of derivatives. Consequently, an iid Gaussian noise additive to the states $\boldsymbol{z}_i(t)$ can translate into correlated non-Gaussian effective noise for $\xi_i(t)$.

A system identification problem can be transformed into parameters estimation problem (or inverse problem) in the form of

$$\boldsymbol{f}^{(i)} = \Phi\boldsymbol{a}^{(i)} + \boldsymbol{\xi}^{(i)}, \tag{19}$$

where $\boldsymbol{f}^{(i)} = [\hat{F}_i(\hat{\boldsymbol{z}}(1)),\dots,\hat{F}_i(\hat{\boldsymbol{z}}(T))]^\top \in \mathbb{R}^{T\times 1}$ represents the estimated function F_i (i-th component of the vector field $\boldsymbol{F}$), $\Phi = [\boldsymbol{\phi}^{(1)},\dots,\boldsymbol{\phi}^{(K)}] \in \mathbb{R}^{T\times K}$ (with $\boldsymbol{\phi}^{(k)} = [\phi_k(\hat{\boldsymbol{z}}(1)),\dots,\phi_k(\hat{\boldsymbol{z}}(T))] \in \mathbb{R}^{T\times 1}$) represent sampled data for the basis functions, $\boldsymbol{\xi}^{(i)} = [\xi_i(1),\dots,\xi_i(T)]^\top \in \mathbb{R}^{T\times 1}$ represents effective noise, and $\boldsymbol{a}^{(i)} = [a_{i1},\dots,a_{iK}]^\top \in \mathbb{R}^{K\times 1}$ is the vector of parameters, which is to be determined. Since the form of the Eq. (19) is the same for each i, we omit the index when discussing the general methodology, and consider the following linear inverse problem:

$$\boldsymbol{f} = \Phi\boldsymbol{a} + \boldsymbol{\xi}, \tag{20}$$

where $\boldsymbol{f} \in \mathbb{R}^{T\times 1}$ and $\Phi \in \mathbb{R}^{T\times K}$ are given, with the goal is to estimate $\boldsymbol{a} \in \mathbb{R}^{K\times 1}$ when the effective noise is not necessarily from independent multivariate Gaussian distribution.

Least squares (LS)

The most commonly used approach to estimate $\boldsymbol{a}$ in Eq. (20) is to use the least squares criterion, which finds $\boldsymbol{a}$ by solving the following least squares minimization problem:

$$\min_{\boldsymbol{a}\in\mathbb{R}^K} \|\Phi\boldsymbol{a} - \boldsymbol{f}\|_2. \tag{21}$$

The solution can be explicitly computed, giving

$$\boldsymbol{a}_{(\mathrm{LS})} = \Phi^\dagger\boldsymbol{f}, \tag{22}$$

where $\Phi^\dagger$ denotes the pseudoinverse of the matrix Φ.[30] Note that in the special case where the minimum is zero (which is unlikely under the presence of noise), the minimizer is not unique and the "least-squares" solution typically refers to a vector $\boldsymbol{a}$ that has the minimal 2-norm and solves the equation $\Phi\boldsymbol{a} = \boldsymbol{f}$. The LS method has several advantages: it is analytically traceable and easy to solve computationally using standard linear algebra routines (e.g., Singular Value Decomposition [SVD]). However, a main disadvantage of the LS approach in system identification, as we discuss in the main text, is that it generally produces a "dense" solution, where most (if not all) components of $\boldsymbol{a}$ are nonzero, which is a severe overfitting of the actual model. This (undesired) feature also makes the method sensitive to noise, especially in the under-sampling regime.

Orthogonal least squares (OLS)

In orthogonal least squares (OLS),[4,31,32] the idea is to iteratively select the columns of Φ that minimize the (2-norm) model error, which corresponds to iterative assigning nonzero values to the components of $\boldsymbol{a}$. In particular, the first step is to select basis $\boldsymbol{\phi}_{k_1}$ and compute the corresponding parameter a_{k_1} and residual $\boldsymbol{r}_1$ according to

$$\begin{cases} (k_1, a_{k_1}) = \arg\min_{k,c}\|\boldsymbol{f} - c\boldsymbol{\phi}_k\|_2, \\ \boldsymbol{r}_1 = \boldsymbol{f} - \boldsymbol{\phi}_{k_1}a_{k_1}. \end{cases} \tag{23}$$

Then, one iteratively selects additional basis functions (until stopping criterion is met) and compute the corresponding parameter value and residual as

$$\begin{cases} (k_{\ell+1}, a_{k_{\ell+1}}) = \arg\min_{k,c}\|\boldsymbol{r}_\ell - c\boldsymbol{\phi}_k\|_2, \\ \boldsymbol{r}_{\ell+1} = \boldsymbol{r}_\ell - \boldsymbol{\phi}_{k_{\ell+1}}a_{k_{\ell+1}}. \end{cases} \tag{24}$$

As for stopping criteria, there are several choices including AIC and Bayesian information criterion (BIC). In this work, in the absence of knowledge of the error distribution, we adopt a commonly used criterion where the iterations terminate when the norm of the residual

Published under license by AIP Publishing.

is below a prescribed threshold. To determine the threshold, we consider 50 log-spaced candidate values in the interval $[10^{-6}, 100]$ and select the best using the 5-fold cross validation.

Lasso

A principled way to impose sparsity on the model structure is to explicitly penalize solution vectors that are nonsparse, by formulating a regularized optimization problem

$$\min_{\boldsymbol{a}\in\mathbb{R}^K} \left(\|\Phi\boldsymbol{a} - \boldsymbol{f}\|_2^2 + \lambda\|\boldsymbol{a}\|_1\right), \tag{25}$$

where the parameter $\lambda \geq 0$ controls the extent to which sparsity is desired: as $\lambda \to \infty$, the second term dominates and the only solution is a vector of all zeros, whereas at the other extreme, $\lambda = 0$ and the problem becomes identical to a least squares problem, which generally yields a full (nonsparse) solution. Values of λ in between then balances the "model fit" quantified by the 2-norm and the sparsity of the solution characterized by the 1-norm. For a given problem, the parameter λ needs to be tuned in order to specify a particular solution. A common way to select λ is via cross validation.[33] In our numerical experiments, we choose λ span according to Ref. 33, with the 5-fold cross validation and 10 values of λ span. We adopt the CVX solver,[34] and from all the solutions found for each λ, we select the solution with minimum residual.

Compressed sensing (CS)

Originally developed in the signal processing literature,[35–37] the idea of compressed sensing (CS) has been adopted in several recent works in the nonlinear system identification.[6,7] Under the CS framework, one solves the following constrained optimization problem:

$$\begin{cases} \arg\min_{\boldsymbol{a}} \|\boldsymbol{a}\|_1, \\ \text{subject to } \|\Phi\boldsymbol{a} - \boldsymbol{f}\| \leq \varepsilon, \end{cases} \tag{26}$$

where the parameter $\varepsilon \geq 0$ is used to relax the otherwise strict constraint $\Phi\boldsymbol{a} = \boldsymbol{f}$, to allow for the presence of noise in data. In our numerical experiments, we choose 10 log-spaced values for $\varepsilon \in [10^{-6}, 100]$, and the 5-fold cross validation. We adopt the CVX solver,[34] and from all the solutions found for each ε, we select the solution with minimum residual.

SINDy

In their recent contribution, Brunton, Proctor, and Kutz introduced SINDy (Sparse Identification of Nonlinear Dynamics) as a way to perform nonlinear system identification.[2] Their main idea is, after formulating the inverse problem (20), to seek a *sparse* solution. In particular, given that Lasso can be computationally costly, they proposed to use sequential least squares with (hard) thresholding as an alternative. For a (prechosen) threshold λ, the method starts from a least squares solution and abandons all basis functions whose corresponding parameter in the solution has absolute value smaller than λ; then, the same is repeated for the data matrix associated with the remaining basis functions, and so on and so forth, until no more basis function (and the corresponding parameter) is removed. For fairness of comparison, we present results of SINDy according to the best threshold parameter λ manually chosen so that no active basis function is removed in the very first step (see KSE example); for the Lorenz system example, we choose $\lambda = 0.02$ as used in a similar example as in Ref. 2.

Tran-Ward (TW)

In their recent paper,[9] Tran and Ward considered the SID problem, where certain fraction of data points are corrupted, and proposed a method to simultaneously identify these corrupted data and reconstruct the system assuming that the corrupted data occur in sparse and isolated time intervals. In addition to an initial guess of the solution and corresponding residual, which can be assigned using standard least squares, the TW approach requires a predetermination of three additional parameters: a tolerance value to set the stopping criterion, threshold value λ used in each iteration to set those parameters whose absolute values are below λ to be zero, and another parameter μ to control the extent to which data points that do not (approximately) satisfy the prescribed model are to be considered as "corrupted data" and removed. For the Lorenz system example, we used the same parameters as in Ref. 9, whereas for the KSE example, we fix $\mu = 0.0125$ (the same used in Ref. 9 and select λ similarly as for the implementation of SINDy.

Implementation details of entropic regression (ER)

As described in the main text, and as shown in details in Algorithm (1), a key quantity to compute in ER is the conditional mutual information $I(X; Y|Z)$ among three (possibly multivariate) random variables X, Y, and Z via samples from these variables, denoted by $(x_t, y_t, z_t)_{t=1,\ldots,T}$. Since the distribution of the variables and their dependences are generally unknown, we adopt a nonparametric estimator for $I(X; Y|Z)$, which is based on statistics of k nearest neighbors.[38] We fix $k = 2$ in all of the reported numerical experiments; we have found that the results change quite minimally when k is varied from this fixed value, suggesting relative robustness of the method.

Another important issue in practice is the determination of threshold under which the conditional mutual information $I(X; Y|Z)$ should be regarded zero. In theory, $I(X; Y|Z)$ is always non-negative and equals zero if and only if X and Y are statistically independent given Z, but such an absolute criterion needs to be softened in practice because the estimated value of $I(X; Y|Z)$ is generally nonzero even when X and Y are indeed independent given Z. A common way to determine whether $I(X; Y|Z) = 0$ or $I(X; Y|Z) > 0$ is to compare the estimated value of $I(X; Y|Z)$ against some threshold. See Sec. (??) for details of robust estimation of the threshold in the context of SID.

SUPPLEMENTARY MATERIAL

See the supplementary material for more details on information theory measurements and additional numerical results for the double-well potential, the Lorenz system, and a coupled network of the logistic map.

ACKNOWLEDGMENTS

This work was funded in part by the Simons Foundation (Grant No. 318812), the Army Research Office (Grant No. W911NF-16-1-0081), the Office of Naval Research (ONR) (Grant No. N00014-15-1-2093), and the DARPA.

Published under license by AIP Publishing.

REFERENCES

[1] N. A. Gershenfeld and A. S. Weigend, *The Future of Time Series. Time Series Prediction: Forecasting the Future and Understanding the Past* (Addison-Wesley, 1993).

[2] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," Proc. Natl. Acad. Sci. U.S.A. **113**, 3932–3937 (2016).

[3] C. Yao and E. M. Bollt, "Modeling and nonlinear parameter estimation with Kronecker product representation for coupled oscillators and spatiotemporal systems," Physica D **1**, 78–99 (2007).

[4] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," Int. J. Control **50**, 1873–1896 (1989).

[5] N. Kalouptsidis, G. Mileounis, B. Babadi, and V. Tarokh, "Adaptive algorithms for sparse system identification," Signal Process. **91**, 1910–1919 (2011).

[6] W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, and C. Grebogi, "Predicting catastrophes in nonlinear dynamical systems by compressive sensing," Phys. Rev. Lett. **106**, 154101 (2011).

[7] W.-X. Wang, Y.-C. Lai, and C. Grebogi, "Data based identification and prediction of nonlinear and complex dynamical systems," Phys. Rep. **644**, 1–76 (2016).

[8] L.-Z. Guo, S. A. Billings, and D. Zhu, "An extended orthogonal forward regression algorithm for system identification using entropy," Int. J. Control **81**, 690–699 (2008).

[9] G. Tran and R. Ward, "Exact recovery of chaotic systems from highly corrupted data," Multiscale Model. Simul. **15**, 1108–1129 (2017).

[10] L. Ljung, "System identification," Wiley Encyclopedia Electrical Electron. Eng. 1–19 (1999).

[11] H. Akaike, "A new look at the statistical model identification," IEEE. Trans. Automat. Contr. **19**, 716–723 (1974).

[12] G. Prando, A. Chiuso, and G. Pillonetto, "Maximum entropy vector kernels for MIMO system identification," Automatica **79**, 326–339 (2017).

[13] J. Sun, C. Cafaro, and E. M. Bollt, "Identifying the coupling structure in complex systems through the optimal causation entropy principle," Entropy **16**, 3416–3433 (2014).

[14] J. Sun, D. Taylor, and E. Bollt, "Causal network inference by optimal causation entropy," SIAM J. Appl. Dyn. Syst. **14**, 73–106 (2015).

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, Hoboken, NJ, 2005), pp. 1–748.

[16] B. Saltzman, "Finite amplitude free convection as an initial value problem—I," J. Atmos. Sci. **19**, 329–341 (1962).

[17] E. N. Lorenz, "Deterministic nonperiodic flow," J. Atmos. Sci. **20**, 130–141 (1963).

[18] Y. Kuramoto and T. Tsuzuki, "Persistent propagation of concentration waves in dissipative media far from thermal equilibrium," Progress Theor. Phys. **55**, 356–369 (1976).

[19] Y. Kuramoto, "Diffusion-induced chaos in reaction systems," Progress Theor. Phys. Suppl. **64**, 346–367 (1978).

[20] G. I. Sivashinsky, "Nonlinear analysis of hydrodynamic instability in laminar flames—I. Derivation of basic equations," Acta Astronaut. **4**, 1177–1206 (1977).

[21] J. M. Hyman and B. Nicolaenko, "The Kuramoto-Sivashinsky equation: A bridge between PDE's and dynamical systems," Physica D **18**, 113–126 (1986).

[22] Y. Lan and P. Cvitanović, "Unstable recurrent patterns in Kuramoto-Sivashinsky dynamics," Phys. Rev. E **78**, 026208 (2008).

[23] F. Christiansen, P. Cvitanovic, and V. Putkaradze, "Spatiotemporal chaos in terms of unstable recurrent patterns," Nonlinearity **10**, 55 (1997).

[24] P. Hohenberg and B. I. Shraiman, "Chaotic behavior of an extended system," Physica D **37**, 109–115 (1989).

[25] P. Cvitanovic, R. Artuso, R. Mainieri, G. Tanner, G. Vattay, N. Whelan, and A. Wirzba, *Chaos: Classical and Quantum* (Niels Bohr Institute, Copenhagen, 2005), Vol. 69.

[26] J. C. Robinson, *Infinite-Dimensional Dynamical Systems: An Introduction to Dissipative Parabolic PDEs and the Theory of Global Attractors*, Cambridge Texts in Applied Mathematics (Cambridge University Press, 2001), ISBN: 9780521635646.

[27] M. S. Jolly, I. Kevrekidis, and E. S. Titi, "Approximate inertial manifolds for the Kuramoto-Sivashinsky equation: Analysis and computations," Physica D **44**, 38–60 (1990).

[28] S. Ramdani, B. Rossetto, L. O. Chua, and R. Lozi, "Slow manifolds of some chaotic systems with applications to laser systems," Int. J. Bifurcat. Chaos **10**, 2729–2744 (2000).

[29] M. S. Jolly, R. M. S. Rosa, and R. M. Temam, "Accurate computations on inertial manifolds," J. Sci. Comp. **22**(6), 2216–2238 (2001).

[30] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. (Johns Hopkins University Press, Baltimore, MD, 2013), ISBN: 1-4214-0859-7.

[31] L. Wang and R. Langari, "Building Sugeno-type models using fuzzy discretization and orthogonal parameter estimation techniques," IEEE Trans. Fuzzy Syst. **3**, 454–458 (1995).

[32] M. Korenberg, S. A. Billings, H. Liu, and P. McIlroy, "Orthogonal parameter estimation algorithm for non-linear stochastic systems," Int. J. Control **48**, 193–210 (1988).

[33] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations* (CRC Press, 2015).

[34] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming," 2008.

[35] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," Commun. Pure Appl. Math. **59**, 1207–1223 (2006).

[36] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," IEEE Trans. Inf. Theory **52**, 489–509 (2006).

[37] D. L. Donoho, "Compressed sensing," IEEE Trans. Inf. Theory **52**, 1289–1306 (2006).

[38] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," Phys. Rev. E **69**, 066–138 (2004).

Published under license by AIP Publishing.