# How Entropic Regression Beats the Outliers Problem in Nonlinear System Identification

Abd AlRahman R. AlMomani[1,5], Jie Sun[1,2,3,4,*], and Erik Bollt[1,2,4,5,*]

[1]Clarkson Center for Complex Systems Science ($C^3S^2$), Potsdam, NY, USA
[2]Department of Mathematics, Clarkson University, Potsdam, NY, USA
[3]Department of Computer Science, Clarkson University, Potsdam, NY, USA
[4]Department of Physics, Clarkson University, Potsdam, NY, USA
[5]Department of Electrical and Computer Engineering, Clarkson University, Potsdam, NY, USA
[*]Correspondence author(s): sunj@clarkson.edu, bolltem@clarkson.edu

August 19, 2018

**Abstract**

System identification (SID) is central in science and engineering applications whereby a general model form is assumed, but active terms and parameters must be inferred from observations. Virtually all methods for SI rely on optimizing some metric-based cost function that describes how a model fits observational data. The most common cost function employs a Euclidean metric and leads to a least squares estimate, while recently it becomes popular to also account for model sparsity such as in compressed sensing and Lasso methods. While the effectiveness of these methods has been demonstrated in various model systems, it remains unclear whether SID can be accomplished under more realistic scenarios where there may be large noise and outliers. We show that sparsity-focused methods such as compressive sensing, when used in the presence of noise, may result in "over sparse" solutions that are brittle to outliers. In fact, metric-based methods are prone to outliers because outliers by nature have an unproportionally large influence. To mitigate such issues of large noise and outliers encountered in practice, we develop an entropic regression approach for nonlinear SID, whereby true model structures are identified based on relevance in reducing information flow uncertainty, not necessarily sparsity. The use of information-theoretic measures as opposed to a metric-based cost function has the unique advantage, due to the asymptotic equipartition property of probability distributions, that outliers and other low-occurrence events are naturally and intrinsically de-emphasized.

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 [15]. 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 to model the delicate but important (physical) effects. 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. Weigenbend 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..." [19].

The problem of learning a (dynamical) system from observational data is commonly known as *system identification* (SID), and usually 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 $m, \gamma$, 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. Filtering methods can be thought of closely related, Kalman filtering [31, 50, 51] being the best-known and it can be extended to models consisting of linear combinations of a general basis terms. 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 is previously known about the system, in a black box manner.

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

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

where $\boldsymbol{z} = [z_1, \ldots, z_N]^T \in \mathbb{R}^N$ and $\boldsymbol{F} = [F_1, \ldots, F_N]^\top : \mathbb{R}^N \to \mathbb{R}^N$. Each component function $F_i(\boldsymbol{z})$ can be represented using a power series (for example a Taylor series or a Fourier series), writing 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. 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 distinguishes the various methods we discuss here. This was the theme in some of our previous work based on a least squares metric [51], as well as the works of others by compressed sensing [2, 6, 7, 8, 9, 17, 44, 45, 48], and recently including SINDy as a general learning method premised again on similar principles [4]. There are conflicting interests however with regard to how many terms to include in the truncated series representation. While a large set of basis functions allows for a rich class of behaviors, too large a set causes problems with numerics, and convergence of fitting accuracy and overfitting, which also requires ever more data to fit an exponential explosion of terms. Recent breakthroughs in nonlinear SID has found a way to overcome this seemingly paradox, by allowing a large set of basis functions and meanwhile imposing sparsity in the model, thus mitigating the issue of overfitting [16]. The success of such sparsity-based SID has been demonstrated in several recent works and applications [11, 18, 24, 28, 46].

Regardless of the particular method or system, most previous work focuses on observational data that are either perfectly sampled from a known system or with some very low level of noise. In practice, since an observation process can subject to large disturbances in unpredictable ways, the effective noise can sometimes be large and even contain "outliers" that may contaminate the otherwise excellent 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. Here we found that relatively large noise and outliers in the observational data generally cause issues for standard SID methods, including those that specialize in finding sparse models. More alarmingly, while the underlying true model may indeed be sparse with respect to a chosen basis, the truth may well not be optimally sparse and consequently sparse optimization methods (such as compressive sensing) can be brittle in the sense that when they fail, they may fail spectacularly, and the presence of outliers makes this issue even more pronounced.

In this work we depart from the standard approaches to 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 will causes the model to put (much) less weights on the "good" data points. Instead, we propose

to infer the (sparsity) structure of a general model together with its parameters using a novel *information theoretic* approach that we call entropic regression because of the inclusion of both entropy optimization and regression. As we will show, while standard metric-based methods emphasize the data in ways as designed by the chosen metric, the proposed entropic regression is robust with regards 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, entropic regression is emphasizing "relevance" according to a model-free, information-theoretic criterion. Basis terms will be included in the model only because they are relevant and not because they together make up a sparse model. We demonstrate the effectiveness of entropic regression in several examples, including Lorenz system, Kuramoto-Sivashinsky equations, and a double well potential, where in each case the observed data contains large noise and outliers. We also remark on the computational complexity and convergence in small-data regime, as well as open problems.

# Results

### Nonlinear System Identification: Problem Statement and Formulation

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. There is not much requirement on the basis functions $\{\phi_k(\boldsymbol{z})\}$, and a common choice is the standard *polynomial basis*

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

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 are 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 $F_i(\boldsymbol{z}(t_k)) = \frac{z_i(t_{k+1}) - z_i(t_k)}{\tau_k} + \mathcal{O}(\tau_k)$ with $\tau_k = t_{k+1} - t_k$. 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 as estimating all the parameters $\{a_{ik}\}$. In practice, the infinite series is truncated after a finite number of terms; such truncation together with observational noise defines a forward model of the inverse problem, as

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

Here $\hat{F}_i$ denotes the empirically observed (noisy) value of $F_i$ and $\xi_i(t)$ represents the aggregated effect of truncation error and additional noise at time $t$. The same can be written as

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

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

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

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

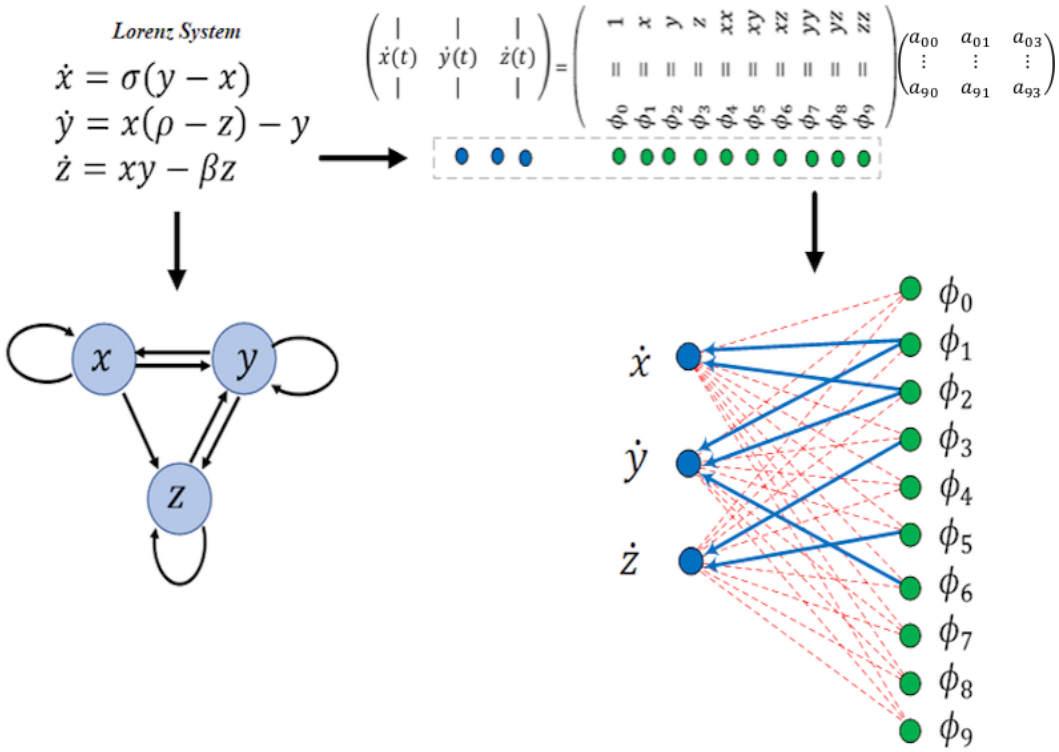


Figure 1: (Left) Lorenz system as a dynamical system and its standard graph representation. See [40]. (Right) Linear combination of nonlinear basis functions, with coupling coefficients $\boldsymbol{a}_{ij}$ forming an influence matrix $\boldsymbol{a}$, or equivalently a coupling graph (right-bottom). Here the directed edges describe the influence of basis terms on the individual variables of the system.

where $\boldsymbol{f}^{(i)} = [\hat{F}_i(\boldsymbol{z}(t_1)), \ldots, \hat{F}_i(\boldsymbol{z}(t_\ell))]^\top \in \mathbb{R}^{\ell \times 1}$ represents sampled data of the function $F_i$ ($i$-th component of the vector field $\boldsymbol{F}$), $\Phi = [\boldsymbol{\phi}^{(1)}, \ldots, \boldsymbol{\phi}^{(K)}] \in \mathbb{R}^{\ell \times K}$ (with $\boldsymbol{\phi}^{(k)} = [\phi_k(\boldsymbol{z}(t_1)), \ldots, \phi_k(\boldsymbol{z}(t_\ell))] \in \mathbb{R}^{\ell \times 1}$) represent sampled data for the basis functions, $\boldsymbol{\xi}^{(i)} = [\xi_i(t_1), \ldots, \xi_i(t_\ell)]^\top \in \mathbb{R}^{\ell \times 1}$ represents noise, and $\boldsymbol{a}^{(i)} = [a_{i1}, \ldots, a_{iK}]^\top \in \mathbb{R}^{K \times 1}$ is the vector of parameters which is to be determined. Since the form of the equation (6) is the same for each $i$, we omit the index when discussing the general methodology, and consider the following linear inverse problem

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

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, lasso, compressed sending, to name a few. Each of these being 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. 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 [43, 42]; in each iteration, the corresponding parameters are updated using a standard regression method, e.g., least squares. 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 below.

The ER method contains two stages (also see Algorithm 1 for the pseudocode for the algorithm): 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). 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

$$\begin{cases} k_1 = \arg\max_k I(\boldsymbol{\phi}_k; \boldsymbol{f}), \\ a_{k_1} = \arg\min_c \|\boldsymbol{f} - c\boldsymbol{\phi}_{k_1}\|_2, \\ \boldsymbol{z}_1 = \boldsymbol{\phi}_{k_1} a_{k_1}. \end{cases} \tag{8}$$

Next, in each iteration of the forward stage, we perform the following computations and updates, for $i = 2, 3, \ldots,$

$$\begin{cases} k_i = \arg\max_k I(\boldsymbol{\phi}_k; \boldsymbol{f}|\boldsymbol{z}_{i-1}), \\ [a_{k_1}, \ldots, a_{k_i}] = \arg\min_{c_1,\ldots,c_i} \|\boldsymbol{f} - c_1\boldsymbol{\phi}_{k_1} - \cdots - c_i\boldsymbol{\phi}_{k_i}\|_2, \\ \boldsymbol{z}_i = a_{k_1}\boldsymbol{\phi}_{k_1} + \cdots + a_{k_i}\boldsymbol{\phi}_{k_i}. \end{cases} \tag{9}$$

The process terminates when either all basis functions are exhausted, or $\max_k I(\boldsymbol{\phi}_k; \boldsymbol{f}|\boldsymbol{z}_{i-1}) = 0$ 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 $S = \{k_1, \ldots, k_m\}$ together with the corresponding parameters $a_{k_1}, \ldots, a_{k_m}$ ($a_j = 0$ 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 these 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

$$\begin{cases} [a_{k_1}, \ldots, a_{k_{i-1}}, a_{k_{i+1}}, \ldots, a_{k_m}] = \arg\min_{c_1,\ldots,c_j} \|\boldsymbol{f} - c_1\boldsymbol{\phi}_{k_1} - \cdots - c_{i-1}\boldsymbol{\phi}_{k_{i-1}} - c_{i+1}\boldsymbol{\phi}_{k_{i+1}} - \cdots - \boldsymbol{\phi}_{k_m}\|_2, \\ \bar{\boldsymbol{z}}_j = a_{k_1}\boldsymbol{\phi}_{k_1} + \cdots + a_{k_{i-1}}\boldsymbol{\phi}_{k_{i-1}} + a_{k_{i+1}}\boldsymbol{\phi}_{k_{i+1}} + a_{k_m}\boldsymbol{\phi}_{k_m}, \end{cases} \tag{10}$$

and updating $S \to S - \{j\}$ if $I(\boldsymbol{\phi}_j; \boldsymbol{f}|\bar{\boldsymbol{z}}_j) = 0$. The result of the backward ER is the reduced set of indices $S = \{\ell_1, \ldots, \ell_n\}$ with $n \leq m$, together with the corresponding parameters $a_{\ell_1}, \ldots, a_{\ell_n}$ ($a_j = 0$ for $j \notin S$) and model $\boldsymbol{f} \approx a_{\ell_1}\boldsymbol{\phi}_{\ell_1} + \cdots + a_{\ell_n}\boldsymbol{\phi}_{\ell_n}$.

In practice, the 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 some significance testing, the details of which are included in the Methods section as well as in *(Supplementary Sec. 3)* with the code.

### Numerical Experiments: Outliers, Expansion Order, and the Paradox of Sparsity

To demonstrate the utility of ER for nonlinear system identification under noisy observations, we compare its performance against common existing methods including the standard least squares (LS), orthogonal least squares (OLS), Lasso, and compressed sensing (CS). 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 ODEs and PDEs. In addition, we consider different noise models and especially the presence of outliers in order to evaluate the robustness of the respective methods.

For each example system, we sample the state of each variable at a uniform rate of $\Delta t$ to obtain a multivariate time series $\{z_k(t_i)\}_{k=1,\ldots,N;i=1,\ldots,\ell}$; then we add noise to each data point and obtain the noisy

**Algorithm 1** Entropic Regression

**procedure** FORWARD ER:$(\mathbf{f}, \mathbf{\Phi}, \mathbf{tol})$
  $K_f = \emptyset, p = \emptyset, v = \infty$
  **while** $v > tol$ **do**
    $K_f \leftarrow p$
    $I_j^{est} := C_{\Phi_j \to \mathbf{f} | \mathbf{\Phi}_{\mathbf{K_f}}}\left(\mathbf{\Phi}^+_{\mathbf{K_f}} \mathbf{f}\right)$, for all $j = 1, ..., K$ and $j \notin K_f$.
    $v, p := \max_j(I_j)$
  **return** $K_f$
**procedure** BACKWARD ER:$(\mathbf{f}, \mathbf{\Phi}, \mathbf{tol}, \mathbf{K_f})$
  $K_b = K_f, p = \emptyset, v = -\infty$
  **while** $v > tol$ **do**
    $K_b := \{K_b\} - \{p\}$
    $I_j^{est} := C_{\Phi_j \to \mathbf{f} | \mathbf{\Phi}_{\mathbf{K_b - j}}}\left(\mathbf{\Phi}^+_{\mathbf{K_b - j}} \mathbf{f}\right)$, for all $j \in K_b$.
    $v, p := \min_j(I_j)$
  **return** $K_b$
**return** $K = K_b$.

empirical time series denoted by $\{\hat{z}_k(t_i)\}$, where

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

with $\eta_{ki}$ represents noise.

**Example 1. Lorenz system.** Our first detailed example data set was generated by noisy observations from a chaotic Lorenz system. The dynamics of the system is represented by a three-dimensional ODE which is a prototype system as a minimal model for thermal convection, obtained by a low-ordered modal truncation of the Saltzman PDE [39], and for many parameter combinations exhibits chaotic behavior [35]. 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).

First, we compare several nonlinear SID methods in reconstructing the Lorenz system when the observational noise is drawn independently from a Gaussian distribution, $\eta \sim \mathcal{N}(0, \epsilon^2)$. As shown in Fig. 2(a), when the level of observational noise is low ($\epsilon = 10^{-5}$), all methods perform well. In the small-data regime, OLS seems to be the best whereas with abundant data, ER takes over. However, for higher level of noise ($\epsilon = 10^{-2}$), we see from Fig. 2(b) that all methods perform poorly except for the proposed ER method which remains effective.

Next, to explore the performance of SID methods under the presence of outliers, we conduct additional numerical experiments. We fix the sample size to be 1000 to ensure that an ample amount of data is available. The extent to which outliers present is controlled by a single parameter $p$: each observation is subject to an added noise $\eta$, where $\eta \sim \mathcal{N}(0, \epsilon_1^2)$ with probability $1-p$ and $\eta \sim \mathcal{N}(0, \epsilon_1^2 + \epsilon_2^2)$ with probability $p$. Here $\epsilon_1 = 10^{-5}$ (low noise) and $\epsilon_2 = 5$ (very high noise). The results of SID are shown in Fig. 3. When $p = 0$, no outlier is present and all methods perform well; as $p$ starts to increase, it is suddenly very challenging for any method to correctly identify the system because of the presence of outliers due to very large noise. Nevertheless, we found that ER is still able to identify a system that fits the true system, while all other methods fail. As shown in the side patches of Fig. 3, under the presence of outliers, the attractors

reconstructed from CS generally look nothing like the true Lorenz "butterfly" attractor, whereas those from ER are visibly indistinguishable from the true attractor.

Given that a major theme of modern SID is to look for *sparse* representations, and the Lorenz system under standard polynomial basis is indeed sparse, it is worth asking: what are the respective structure identified by the different methods? In Fig. 4 we compare the structure of the identified model using different methods across a range of parameter values for $\rho$. As it turns out, most methods (including for ER) do return relatively sparse structure, but often it is a wrong structure that has a lot of false positives. Alarmingly, the CS method, which typically looks to optimize sparsity, is producing an even sparser solution than the true solution!

The mere fact that LS fails is perhaps not surprising, given that it tends to produce "dense" solutions. A natural question one might ask, given that several of the methods were designed to work for sparse regression, is that whether identification can be (more) successful if the underling problem is indeed sparse. To address this question, we conduct system identification from the same data but under different number of basis functions. As the number of basis functions increases, the problem becomes more sparse and one would typically expect methods such as Lasso and CS to perform better. Interestingly though, these methods not only do not yield better identification outcomes, they actually tend to produce worse results, see *(Supplementary Sec. 4.2)*. The proposed ER, on the other hand, is robust under the varying and increasing number of terms.

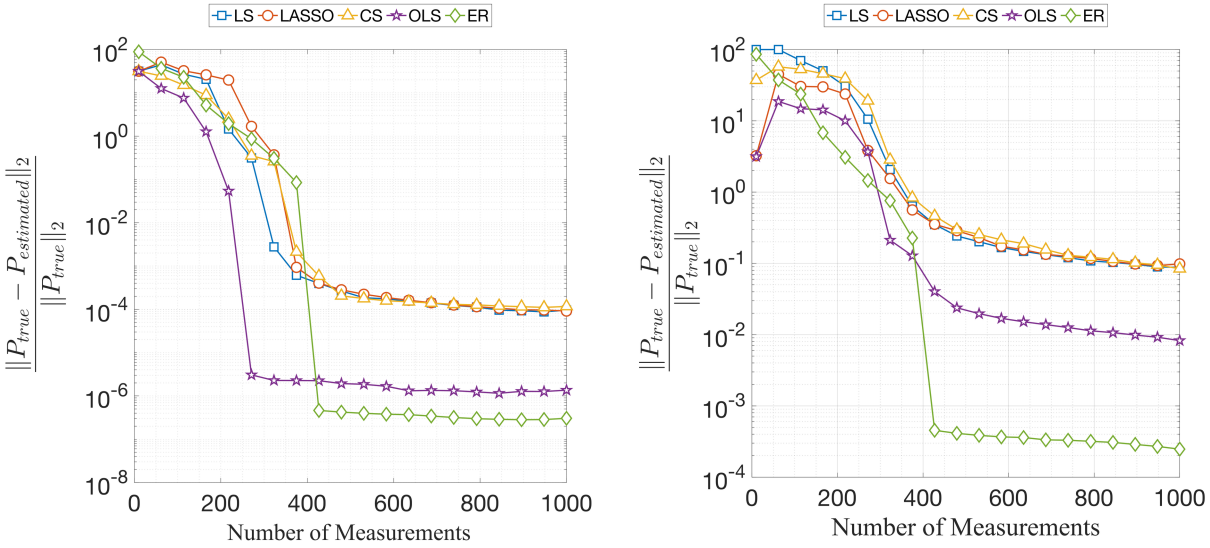


Figure 2: Lorenz system. (Left) Relative error in the parameter estimation for a Lorenz system given in Eq. 12 but subject to noisy measurements by Gaussian noise, $\eta \sim \mathcal{N}(0, \epsilon_1)$, with standard deviation $\epsilon_1 = 10^{-5}$, and using a $5^{th}$-order polynomial expansion. We see that ER has an overall superior performance compared to others standard methods, especially as the number of measurements increases. In this low-noise setting, clearly OLS also performs well whereas LS as expected requires a larger number of measurements to reach reasonable accuracy and CS shows low performance in complex dynamic such as Lorenz, for reasons related "over-sparsing" as discussed in the text. (Right) Again the same fitted Lorenz system and parameters but now with larger noise $\epsilon_1 = 10^{-2}$. Contrast these to outlier "corrupted" versions as shown in Fig. 3.

**Example 2. Kuramoto-Sivashinsky equations.** To further demonstrate the power of ER, we consider a nonlinear PDE, namely the Kuramoto-Sivashinsky (KS) equation [32, 33, 41, 25, 34], 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 [13, 23]. We

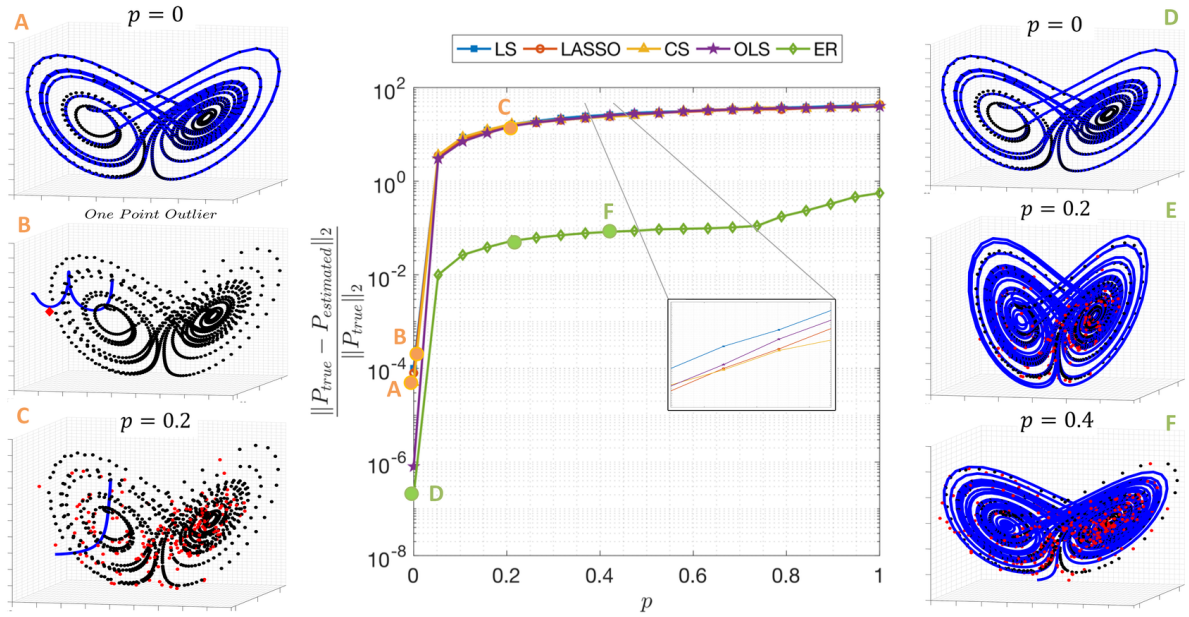


Figure 3: SID for the Lorenz system when the observations are corrupted by outliers. Contrast to Fig. 2. As before, we specify a level of persistent Gaussian observation noise, $\eta \sim \mathcal{N}(0, \epsilon_1)(1 - Ber(p))$, but now furthermore we allow for an "outlier noise", as "occasional" bursts of much larger perturbations, $\eta \sim \mathcal{N}(0, \epsilon_1 + \epsilon_2)Ber(p)$, 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 Lorenz system given in Eq. 12 with noise, $\epsilon_1 = 10^{-5}$, $5^{th}$-Order polynomial expansion, 1000 Measurements, but now the signal is further corrupted by $\epsilon_2 = 5$. We see that only ER has low error while all other solvers have high error indicated a failure in estimation process. Zoom-view at $p = 0.4$ shows that the solvers are not identical, LS has very high error according to its sensitivity to outliers and the CS has the lower value because it finally minimize the $L^1$ norm. But we see clearly that once $0 < p$, the "relative" error of all solvers become large. (Left) In the left column we see the constructed dynamic by the recovered solution of CS. We see that At low base noise $\epsilon_1 = 10^{-5}$ and no corrupted measurements $p = 0$, CS has good results in recovering the dynamic. With one single point outliers (marked in green color middle-left figure) we see that CS completely fails in recovering the dynamic. (Right) In the left column we see the constructed dynamic by the recovered solution of ER. Even with high probability ratio $p$, we see ER is robust to noise and outliers and it recovered the true dynamic.

will consider Kuramoto-Sivashinsky system in the following form,

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

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 $\nu$ controls the suppression of solutions with fast spatial variations, and is set to $\nu = 0.029910$ under which the system exhibit chaotic behavior [13].

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 ODE's in the corresponding Banach space.

To develop the Galerkin projection, we follow the procedure as presented in [1], to expand a periodic solution $u(x, t)$ using a discrete spatial Fourier series,

$$u(x, t) = \sum_{-\infty}^{\infty} b_k(t) e^{ikqx}, \qquad \text{where } q = \frac{2\pi}{L}. \tag{13}$$

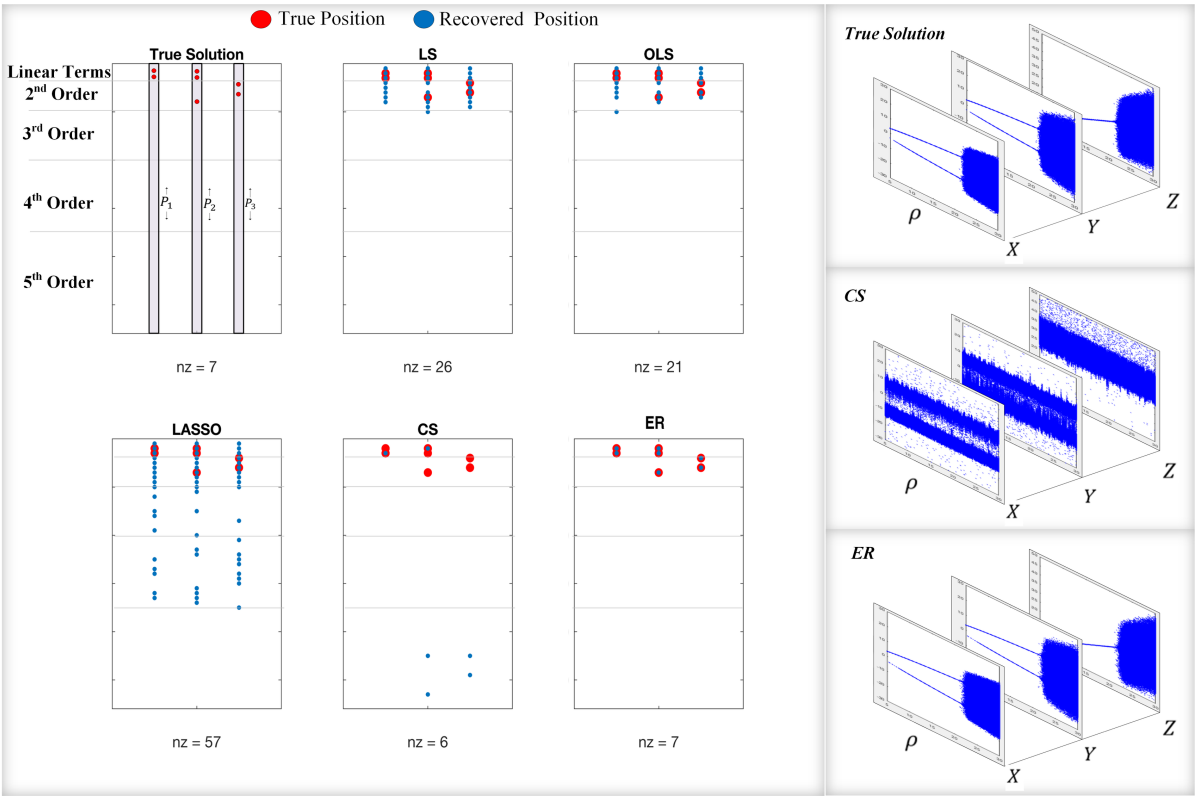


Figure 4: Sparse representation of the solution found by solvers using 1000 measurements, $\epsilon_1 = 10^{-5}$, $\epsilon_2 = 5$ and $p = 0.2$. The upper left corner shows the true solution of Lorenz system with details to improve the readability of other figures. The right column shows the bifurcation diagram with $\rho \in [5, 30]$ as bifurcation parameter, created using 5000 initial conditions evolved according the recovered solution.

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 [37, 38] with the representation of a PDE as infinitely many ODE's 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 pure imaginary, $b_k = ia_k$ and $a_k \in \mathbb{R}$. Then,

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

where $k = 1, .., N_m$. However, the assumption that there is a slow manifold [37, 3] (slow modes as an inertial manifold [38, 37, 26, 36, 27]) suggest the practical matter that a finite truncation of the series Eq. (13), and correspondingly the a 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, have meaning to estimate the structure and parameters of the ODE model (15).

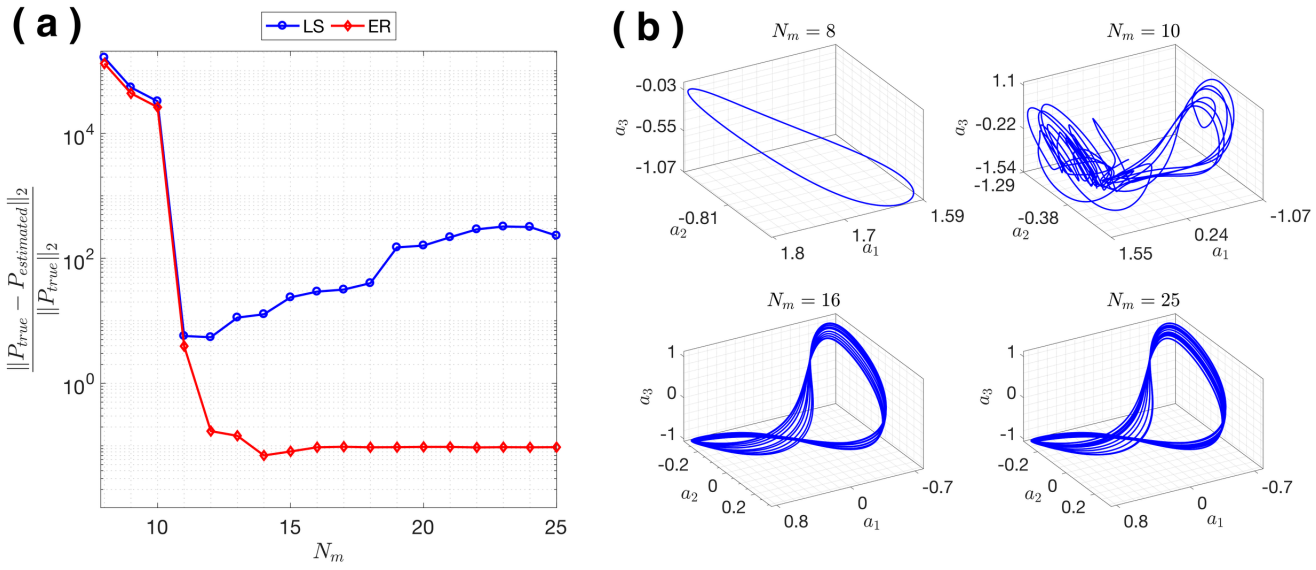


Figure 5: **(a)** Relative error in estimated parameters using 5000 measurements and noise level $\epsilon_1 = 0.001$. We see that from Eq. 15, the linear terms grow to high values as $N_m$ increase, and because of the sloppiness of KSE parameters and the hight magnitude of the sloppy parameters, we limit the error computation the first 100 terms. **(b)** We show the first three modes $a_1, a_2$ and $a_3$ for selected number of modes. 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).

Fig. 5-(a) shows the relative error in estimated parameters for KSE under different number of modes for LS and ER. Fig. 5-(b) shows the first three dimensions plot under different number of modes. We see that using few number of modes ($N_m = 8$ and $N_m = 10$) leads to a failure in constructing the true dynamic that describe the system, and using large number of modes ($N_m = 25$) may be unnecessary since the dynamic can be described with lower number ($N_m = 16$) with no significant extra information added by increasing $N_m$.

Fig. 6 shows the relative error in estimating parameters with different number of modes $N_m$, and the constructed dynamic for selected number of modes. We compute the error for LS and ER only and excluded other solvers for this experiment because of the computation complexity of other solvers as discussed before.

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. For more details for this example see *(Supplementary Sec. 4.4)*.

**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 $\Phi$ using the $10^th$ order polynomial expansion with $K = 11$ is the number of candidate functions. Then, we consider a single fixed value corrupted measurement to be $f(0.6) = 0.2$.

Fig. 7 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

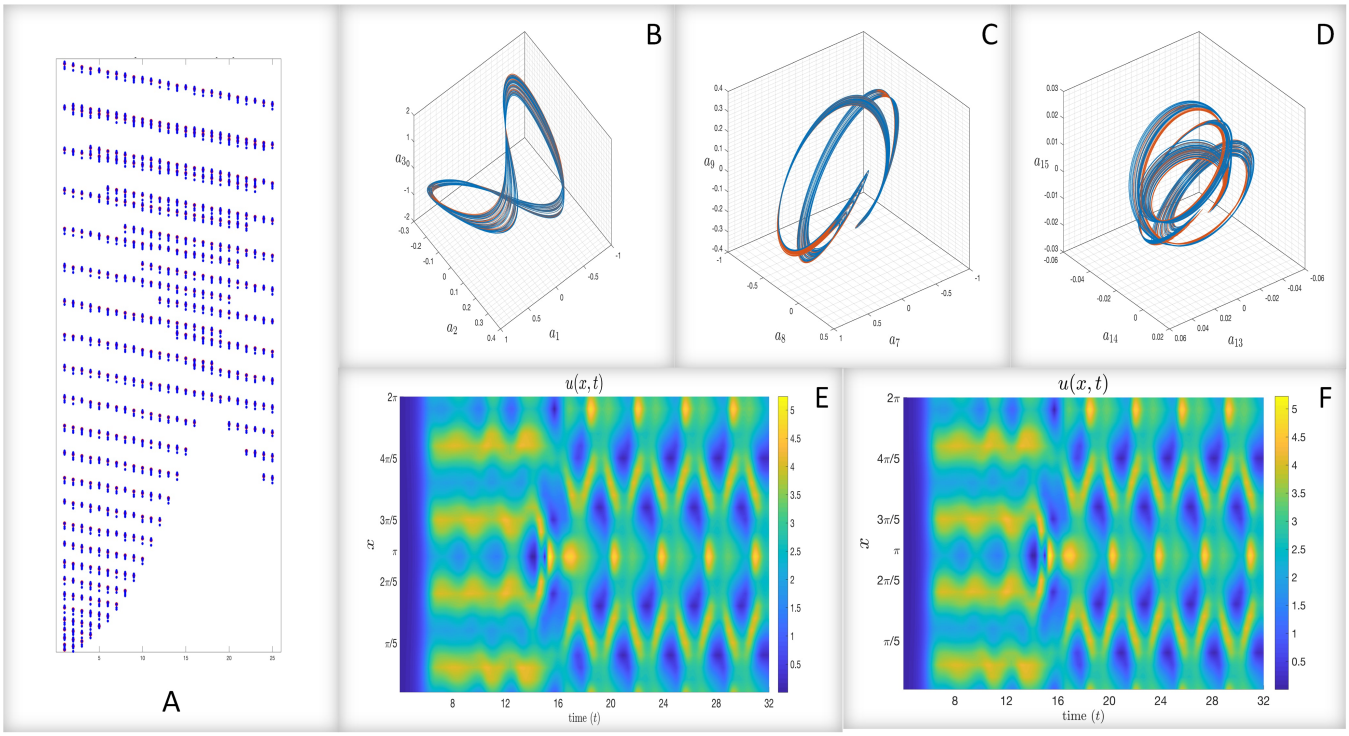


Figure 6: (A) In analogy to Fig. 4, sparse representation of ER solution. (B-D) 3-D plot of selected modes. Blue line represent the dynamic constructed by the true solution, and the Red line represent the dynamic constructed with ER solution. (E-F) $u(x,t)$ constructed by the true solution and ER solution respectively. This shows that even we have false positive positions and relative error 0.1, but ER detected the parameters that have the true influence to the dynamic and large scale view of $u(x,t)$ solution shows that ER recover the underlying dynamic accurately.

sake of clearness, Fig. 7 shows the results for CS and ER. The results for each solver and details are provided in *(Supplementary Sec. 4.1)* in addition to more numerical examples.

## Discussion

The main theme of the paper is on nonlinear system identification (SID) under potentially large noise and outliers, which is to learn the functional form and parameters of a nonlinear system based on observations of its states. 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. As we demonstrate using several example systems, including the chaotic Lorenz system and the Kuramoto-Sivashinsky equations, that existing SID methods are not robust against noise, and can be quite sensitive to the presence of outliers. We identify the root cause of such non-robustness being 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

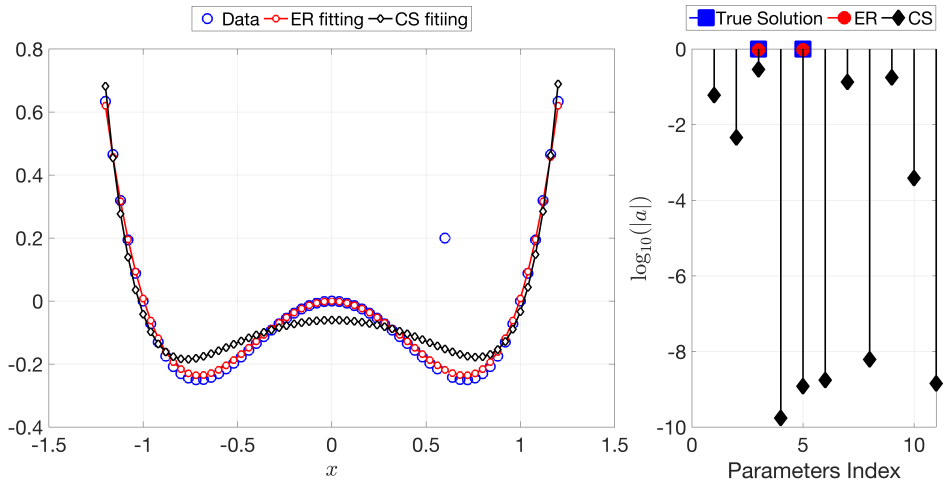


Figure 7: Double well potential given by Eq. (16) data fitting using ER and CS. CS solution found as the solution with minimum residual from 100 log-spaced values of $\epsilon \in [10^{-9}, 10^{2}]$.

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, and in this sense ER is superior to any other method that we know of for such settings. On a more fundamental level, ER's robustness against outliers can be attributed to an important principle in information theory called the asymptotic equipartition property (AEP) [14]. The outcome of this principle is that sampled data can be partitioned into "typical" samples and "atypical" samples, with the rare atypical samples end 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 interpretion 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 data samples of the variables $\boldsymbol{z}$ and the vector field $F_i$, 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 $t$-th sample satisfies the equation

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

Here $\xi_i(t)$ represents the accumulative effects of truncation error and (additional) observational noise.

Having transformed a system identification problem into an parameter estimation problem (or inverse problem) in the form of

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

where $\boldsymbol{f}^{(i)} = [\hat{F}_i(\boldsymbol{z}(1)), \ldots, \hat{F}_i(\boldsymbol{z}(T))]^\top \in \mathbb{R}^{T\times 1}$ represents sampled data of the function $F_i$ ($i$-th component of the vector field $\boldsymbol{F}$), $\Phi = [\boldsymbol{\phi}^{(1)}, \ldots, \boldsymbol{\phi}^{(K)}] \in \mathbb{R}^{T\times K}$ (with $\boldsymbol{\phi}^{(k)} = [\phi_k(\boldsymbol{z}(1)), \ldots, \phi_k(\boldsymbol{z}(T))] \in \mathbb{R}^{T\times 1}$) represent sampled data for the basis functions, $\boldsymbol{\xi}^{(i)} = [\xi_i(1), \ldots, \xi_i(T)]^\top \in \mathbb{R}^{T\times 1}$ represents noise, and $\boldsymbol{a}^{(i)} = [a_{i1}, \ldots, a_{iK}]^\top \in \mathbb{R}^{K\times 1}$ is the vector of parameters which is to be determined. Since the form of the equation (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 $\boldsymbol{a} \in \mathbb{R}^{K\times 1}$.

### 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 $\Phi$ [20]. 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., 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 "full" solution, where each component of $\boldsymbol{a}$ is nonzero despite the actual system structure typically being sparse. 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) [10, 47, 29], the idea is to iteratively select the columns of $\Phi$ that minimize the 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 further basis and computing the corresponding parameter value and residual, as

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

A simple stopping criterion is when the norm of the residual is below a certain threshold; although one can also apply more sophisticated criteria such as AIC and BIC. In this work, we considered the simple stopping criterion, with 10 log-spaced threshold values $\in [10^{-6}, 1]$, and 5-Folds cross validation.

### Lasso

Another way to impose sparsity on the model structure is to explicitly penalize solution vectors that are non-sparse, 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 (non-sparse) solution. Values of $\lambda$ 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, each $\lambda$ gives a solution, and thus the parameter $\lambda$ needs to be selected in order to properly define a solution. A common way to select $\lambda$ is via cross validation [22]. In our numerical experiments, we choose $\lambda$ span according to [22], with 5-Folds cross validation and 10 values $\lambda$ span. We adopt the CVX solver [21], and from all the solutions found for each $\lambda$ we select the solution with minimum residual.

### Compressed sensing (CS)

Originally developed in the signal processing literature, the idea of compressed sensing (CS) has been adopted in several recent work in nonlinear system identification [12, 49, 4, 5] Under the CS framework, one solves the following constrained optimization problem,

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

where the parameter $\epsilon \geq 0$ 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 $\epsilon \in [10^{-6}, 1]$, and 5-Folds cross validation. We adopt the CVX solver [21], and from all the solutions found for each $\epsilon$ we select the solution with minimum residual.

## 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 [30]. 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 nonnegative and equals zero if and only if $X$ and $Y$ are statistically independent given $Z$, but such 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 *(Supplementary Sec.3)* for details of robust estimation of the threshold in the context of SID and also code.

# References


[1] Roberto Artuso, Freddy Christiansen, Ronnie Mainieri, Henrik Hans Henrik Rugh, Gregor Tanner, Niall Whelan, Andreas Wirzba, Predrag Cvitanovi, Artuso Freddy, Christiansen Per, Dahlqvist Ronnie, Mainieri Hans, Henrik Hans Henrik Rugh, Gregor Tanner, Niall Whelan, and Andreas Wirzba. Classical and Quantum Chaos. *Chaos*, 2002.

[2] R G Baraniuk. Compressive Sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, 2007.

[3] Erik Bollt. Attractor Modeling and Empirical Nonlinear Model Reduction of Dissipative Dynamical Systems. *International Journal Of Bifurcation And Chaos*, 17(4):1199–1219, 2007.

[4] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data: Sparse identification of nonlinear dynamical systems. *arXiv*, 1(609):1–26, 2015.

[5] Steven L. Brunton, Joshua L. Proctor, Jonathan H. Tu, and Nathan Kutz. Compressed sensing and dynamic mode decomposition. *Journal of Computational Dynamics*, 2(2158 2491 2015 2 165):165, 2015.

[6] E.J. Candes and M.B. Wakin. An Introduction To Compressive Sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[7] Emmanuel J. Candès. Compressive sampling. *Proceedings of the International Congress of Mathematicians*, pages 1433–1452, 2006.

[8] Emmanuel J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[9] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[10] S. Chen, S. a. Billings, and W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50(769892610):1873–1896, 1989.

[11] Yilun Chen, Yuantao Gu, and Alfred O. Hero. Sparse LMS for system identification. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2009.

[12] Yu-Zhong Chen and Ying-Cheng Lai. Sparse dynamical boltzmann machine for reconstructing complex networks with binary dynamics. *Phys. Rev. E*, 97:032317, Mar 2018.

[13] F. Christiansen, P. Cvitanovi??, and V. Putkaradze. Spatiotemporal chaos in terms of unstable recurrent patterns. *Nonlinearity*, 1997.

[14] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2005.

[15] James P Crutchfield and Bruce S McNamara. Equations of motion from a data series. *Complex Systems*, pages 1–35, 1987.

[16] Tom Dietterich. Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, 1995.

[17] D L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[18] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, 2003.

[19] Neil A. Gershenfeld, Andreas S. Weigend, N. A. Gershenfeld, and Andreas S. Weigend. The future of time series. *Time Series Prediction: Forecasting the Future and Understanding the Past*, 1993.

[20] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (4th Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 2013.

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

[22] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. *Crc*, 2015.

[23] P. C. Hohenberg and Boris I. Shraiman. Chaotic behavior of an extended system. *Physica D: Nonlinear Phenomena*, 1989.

[24] Yi Hao Hsiao, Chaur Chin Chen, Sun In Lin, and Fang Pang Lin. Real-world underwater fish recognition and identification, using sparse representation. *Ecological Informatics*, 2014.

[25] James M. Hyman and Basil Nicolaenko. The Kuramoto-Sivashinsky equation: A bridge between PDE'S and dynamical systems. *Physica D: Nonlinear Phenomena*, 1986.

[26] M. S. Jolly, I. G. Kevrekidis, and E. S. Titi. Approximate inertial manifolds for the Kuramoto-Sivashinsky equation: Analysis and computations. *Physica D: Nonlinear Phenomena*, 1990.

[27] M. S. Jolly, R Rosa, and R Temam. Accurate Computations on Inertial Manifolds, 2001.

[28] Nicholas Kalouptsidis, Gerasimos Mileounis, Behtash Babadi, and Vahid Tarokh. Adaptive algorithms for sparse system identification. *Signal Processing*, 2011.

[29] M. Korenberg, S. A. Billings, Y. P. Liu, and P. J. McIlroy. Orthogonal parameter estimation algorithm for non-linear stochastic systems. *International Journal of Control*, 1988.

[30] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 2004.

[31] Arthur J Krener. The Local Convergence of the Extended Kalman Filter. *xx*, xx:1–13, 2002.

[32] Y. Kuramoto and T. Tsuzuki. Persistent Propagation of Concentration Waves in Dissipative Media Far from Thermal Equilibrium. *Progress of Theoretical Physics*, 1976.

[33] Yoshiki Kuramoto. Diffusion-Induced Chaos in Reaction Systems. *Progress of Theoretical Physics Supplement*, 1978.

[34] Yueheng Lan and Predrag Cvitanović. Unstable recurrent patterns in Kuramoto-Sivashinsky dynamics. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 2008.

[35] Edward N. Lorenz. Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 1963.

[36] S. Ramdani, B. Rossetto, L.O. Chua, and R. Lozi. Slow manifolds of some chaotic systems with applications to laser systems. *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, 2000.

[37] James C. Robinson. Inertial manifolds for the kuramoto-sivashinsky equation. *Physics Letters A*, 184(2):190 – 193, 1994.

[38] 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.

[39] Barry Saltzman. Finite Amplitude Free Convection as an Initial Value ProblemI. *Journal of the Atmospheric Sciences*, 1962.

[40] Hiroki Sayama. *Introduction to the Modeling and Analysis of Complex Systems*. SUNY Binghamton. SUNY Open Textbooks, 2015.

[41] G. I. Sivashinsky. Nonlinear analysis of hydrodynamic instability in laminar flames-I. Derivation of basic equations. *Acta Astronautica*, 1977.

[42] J. Sun, D. Taylor, and E. Bollt. Causal network inference by optimal causation entropy. *SIAM Journal on Applied Dynamical Systems*, 14(1):73–106, 2015.

[43] Jie Sun, Carlo Cafaro, and Erik M. Bollt. Identifying the coupling structure in complex systems through the optimal causation entropy principle. *Entropy*, 2014.

[44] S. Vaegler, D. Stsepankou, J. Hesser, and O. Sauer. SUD11601: A Novel Reconstuction Framework of Prior Image Constrained Compressed Sensing (PICCS) Enabling the Use of Prior Images with Major Deviations. In *Medical Physics*, 2013.

[45] Sven Vaegler, Dzmitry Stsepankou, Jürgen Hesser, and Otto Sauer. Incorporation of local dependent reliability information into the Prior Image Constrained Compressed Sensing (PICCS) reconstruction algorithm. *Zeitschrift fur Medizinische Physik*, 2015.

[46] Palghat P. Vaidyanathan and Piya Pal. System identification with sparse coprime sensing. *IEEE Signal Processing Letters*, 2010.

[47] Liang Wang and Reza Langari. Building Sugeno-Type Models Using Fuzzy Discretization and Orthogonal Parameter Estimation Techniques. *IEEE Transactions on Fuzzy Systems*, 1995.

[48] Wen-Xu Wang, Ying-Cheng Lai, and Celso Grebogi. Data based identification and prediction of nonlinear and complex dynamical systems. *Physics Reports*, 644:1–76, 2016.

[49] Wen Xu Wang, Rui Yang, Ying Cheng Lai, Vassilios Kovanis, and Celso Grebogi. Predicting catastrophes in nonlinear dynamical systems by compressive sensing. *Physical Review Letters*, 106(15), 2011.

[50] Greg Welch and Gary Bishop. An Introduction to the Kalman Filter. *In Practice*, 7(1):1–16, 2006.

[51] Chen Yao and Erik M. Bollt. Modeling and nonlinear parameter estimation with kronecker product representation for coupled oscillators and spatiotemporal systems. *Physica D*, 1(227):78–99, 2007.

## Acknowledgements

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 Grant No. N00014-15-1-2093, and also DARPA.

## Author contributions statement

Jie Sun and Erik Bollt developed the theoretical framework and designed the overall research. Jie Sun developed the prototype algorithm for entropic regression, and Abd AlRahman AlMomani completed its implementation in Matlab. Abd AlRahman AlMomani conducted the numerical experiments. All authors discussed the results and contributed to the writing of the final manuscript.

## Conflict of interest statement

The authors declare no competing financial interests.