MULTISCALE MODEL. SIMUL.
Vol. 19, No. 3, pp. 1474–1497

WEAK SINDy: GALERKIN-BASED DATA-DRIVEN MODEL SELECTION*

DANIEL A. MESSENGER† AND DAVID M. BORTZ†

Abstract. We present a novel weak formulation and discretization for discovering governing equations from noisy measurement data. This method of learning differential equations from data fits into a new class of algorithms that replace pointwise derivative approximations with linear transformations and variance reduction techniques. Compared to the standard SINDy algorithm presented in [S. L. Brunton, J. L. Proctor, and J. N. Kutz, *Proc. Natl. Acad. Sci. USA*, 113 (2016), pp. 3932–3937], our so-called weak SINDy (WSINDy) algorithm allows for reliable model identification from data with large noise (often with ratios greater than 0.1) and reduces the error in the recovered coefficients to enable accurate prediction. Moreover, the coefficient error scales linearly with the noise level, leading to high-accuracy recovery in the low-noise regime. Altogether, WSINDy combines the simplicity and efficiency of the SINDy algorithm with the natural noise reduction of integration, as demonstrated in [H. Schaeffer and S. G. McCalla, *Phys. Rev. E*, 96 (2017), 023302], to arrive at a robust and accurate method of sparse recovery.

Key words. data-driven model selection, nonlinear dynamics, sparse recovery, generalized least squares, Galerkin method, adaptive grid

AMS subject classifications. 37M10, 62J99, 62-07, 65R99

DOI. 10.1137/20M1343166

1. Problem statement.

Consider a first-order dynamical system in D dimensions of the form

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{F}(\boldsymbol{x}(t)), \quad \mathbf{x}(0) = \mathbf{x}_0 \in \mathbb{R}^D, \quad 0 \le t \le T, \tag{1.1}$$

and measurement data $\mathbf{y} \in \mathbb{R}^{M\times D}$ given at M timepoints $\mathbf{t} = (t_1, \dots, t_M)^T$ by

$$\mathbf{y}_{md} = \mathbf{x}_d(t_m) + \epsilon_{md}, \qquad m \in [M],\ d \in [D],$$

where throughout we use the bracket notation $[M] := \{1, \dots, M\}$. The variable $\epsilon \in \mathbb{R}^{M\times D}$ represents a matrix of independent and identically distributed measurement noise. The focus of this article is the reconstruction of the dynamics (1.1) from the measurements $\mathbf{y}$.

The SINDy algorithm (sparse identification of nonlinear dynamics [4]) has been shown to be successful in solving this problem for sparsely represented nonlinear dynamics when noise is small and dynamic scales do not vary across multiple orders of magnitude. This framework assumes that the function $\mathbf{F} : \mathbb{R}^D \to \mathbb{R}^D$ in (1.1) is

*Received by the editors June 8, 2020; accepted for publication (in revised form) June 14, 2021; published electronically September 7, 2021.

https://doi.org/10.1137/20M1343166

Funding: This research was supported in part by the NSF/NIH Joint DMS/NIGMS Mathematical Biology Initiative grant R01GM126559 and in part by the NSF Computing and Communications Foundations Division grant CCF-1815983. This work also utilized resources from the University of Colorado Boulder Research Computing Group, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University.

†Department of Applied Mathematics, University of Colorado, Boulder, CO 80309-0526 USA (daniel.messenger@colorado.edu, dmbortz@colorado.edu).

given componentwise by

$$\mathbf{F}_d(\mathbf{x}(t)) = \sum_{j=1}^{J} \mathbf{w}^\star_{jd} f_j(\mathbf{x}(t)) \tag{1.2}$$

for some known family of functions $(f_j)_{j\in[J]}$ and a sparse weight matrix $\mathbf{w}^\star \in \mathbb{R}^{J\times D}$. The problem is then transformed into solving for $\mathbf{w}^\star$ by building a data matrix $\Theta(\mathbf{y}) \in \mathbb{R}^{M\times J}$ given by

$$\Theta(\mathbf{y})_{mj} = f_j(\mathbf{y}_m), \qquad \mathbf{y}_m := (\mathbf{y}_{m1}, \dots, \mathbf{y}_{mD}),$$

so that the candidate functions are directly evaluated at the noisy data. Solving (1.1) for $\mathbf{F}$ then reduces to identifying a sparse weight matrix $\widehat{\mathbf{w}}$ such that

$$\dot{\mathbf{y}} \approx \Theta(\mathbf{y})\, \widehat{\mathbf{w}}, \tag{1.3}$$

where $\dot{\mathbf{y}}$ is the numerical time derivative of the data $\mathbf{y}$. Sequential-thresholding least squares is then used to arrive at a sparse solution.

1.1. Background. Research into statistically rigorous selection of mathematical models from data can be traced back to Akaike's seminal work in the 1970s [1, 2]. In the last 20 years, there has been substantial work in this area at the interface between applied mathematics, computer science, and statistics (see [3, 11, 12, 19, 22, 23] for both theory and applications). More recently, the formulation of system discovery problems in terms of a candidate basis of nonlinear functions (1.2) and subsequent discretization (1.3) was introduced in [21] in the context of catastrophe prediction. The authors of [21] used compressed sensing techniques to enforce sparsity. Since then there has been an explosion of interest in the problem of identifying nonlinear dynamical systems from data, with some of the primary techniques being Gaussian process regression [15], deep neural networks [16], Bayesian inference [26, 27], and classical methods from numerical analysis [7, 9, 25]. These techniques have been successfully applied to the discovery of both ordinary and partial differential equations.

The variety of discovery algorithms qualitatively differ in the interpretability of the resulting data-driven dynamical system, the scope and efficiency of the algorithm, and the robustness to noise, scale separation, etc. For instance, a neural network based data-driven dynamical system does not easily lend itself to physical interpretation, while the SINDy algorithm identifies governing equations which can be analyzed directly. Moreover, it is also well-known that the training stage for neural networks and other iterative learning algorithms can be computationally costly. Concerning the scope of an algorithm, several methods have been independently developed to discover models under the assumption of some prior knowledge of the governing equations, notably for low-degree polynomial chaotic systems, cyclic ODEs, interacting particles, and Hamiltonian dynamics [20, 18, 13, 24]. In each of these cases the authors derive probabilistic recovery guarantees depending on the number of available trajectories, the size of the candidate model library, the level of incoherence of the data, and/or the sparsity of the governing equations.

The vast majority of algorithms and recovery guarantees assume that pointwise derivatives of the data either are available or can be reliably computed. This severely limits an algorithm's robustness to noise and hence its applicability to real world data. Here we relax this assumption and provide rigorous justification for the weak formulation of the dynamics as a means to circumvent this ubiquitous problem in model

selection. Building off of the SINDy framework, we present the robust discovery algorithm WSINDy (weak SINDy), which operates under the assumption that the time derivative is unavailable and that the only prior knowledge of the governing equations is their inclusion in a large model library. We also focus on the realistic scenario where only a single noisy trajectory of the state variable is available; however, extension to multiple trajectories is of course possible. For simplicity, we restrict numerical experiments to autonomous ODEs for their amenability to analysis. Natural next steps are to explore identification of PDEs and nonautonomous dynamical systems. We note that the use of integral equations for system identification was introduced in [17], where compressed sensing techniques were used to enforce sparsity, and that this technique can be seen as a special case of the method introduced here.

In section 2 we introduce the algorithm with analysis of the resulting error structure. Section 3 contains numerical results showing identification of six ODE systems over a range of noise levels and parameter regimes. In section 4, we provide concluding remarks as well as natural next directions for this line of research. In Appendix A we include a detailed comparison between WSINDy and SINDy as well as further information on the generalized least squares method.

2. WSINDy. We approach the problem of system identification (1.3) from a nonstandard perspective by utilizing the weak form of the differential equation. Recall that for any smooth test function $\phi : \mathbb{R} \to \mathbb{R}$ (absolutely continuous is sufficient) and interval $(a, b) \subset [0, T]$, (1.1) admits the weak formulation

(2.1)
$$\phi(b)\mathbf{x}(b) - \phi(a)\mathbf{x}(a) - \int_a^b \phi'(u)\,\mathbf{x}(u)\,du = \int_a^b \phi(u)\,\mathbf{F}(\mathbf{x}(u))\,du, \qquad 0 \le a < b \le T.$$

With $\phi = 1$, we arrive at the integral equation of the dynamics explored in [17]. If we instead take ϕ to be nonconstant and compactly supported in (a, b), we arrive at

$$-\int_a^b \phi'(u)\,\mathbf{x}(u)\,du = \int_a^b \phi(u)\,\mathbf{F}(\mathbf{x}(u))\,du. \tag{2.2}$$

Assuming a representation of the form (1.2), we then define the generalized residual $\mathcal{R}(\mathbf{w}; \phi)$ for a given test function ϕ by replacing $\mathbf{F}$ with a candidate element from the span of $(f_j)_{j\in[J]}$ and $\mathbf{x}$ with $\mathbf{y}$ as follows:

$$\mathcal{R}(\mathbf{w}; \phi) := \int_a^b \left(\phi'(u)\,\mathbf{y}(u) + \phi(u) \left(\sum_{j=1}^{J} \mathbf{w}_j\, f_j(\mathbf{y}(u)) \right) \right) du. \tag{2.3}$$

Clearly, with $\mathbf{w} = \mathbf{w}^\star$ and $\mathbf{y} = \mathbf{x}(t)$ we have $\mathcal{R}(\mathbf{w}; \phi) = 0$ for all ϕ compactly supported in (a, b); however, $\mathbf{y}$ is a discrete set of data, so (2.3) can at best be approximated numerically. Measurement noise then presents a significant barrier to accurate indentification of $\mathbf{w}^\star$.

2.1. Method overview. For analogy with traditional Galerkin methods, consider the forward problem of solving a dynamical system such as (1.1) for $\mathbf{x}$. The Galerkin approach is to seek a solution $\mathbf{x}$ represented in a chosen trial basis $(f_j)_{j\in[J]}$ such that the residual $\mathcal{R}$, defined by

$$\mathcal{R} = \int \phi(t)(\dot{\mathbf{x}}(t) - \mathbf{F}(\mathbf{x}(t)))\,dt,$$

is minimized over all test functions ϕ living in the span of a given test function basis $(\phi_k)_{k\in[K]}$. If the trial and test function bases are known analytically, inner products of the form $\langle f_j, \phi_k\rangle$ appearing in the residual can be computed exactly. Thus, the computational error results only from representing the solution in a finite-dimensional function space.

The method we present here can be considered a data-driven Galerkin method of solving for $\mathbf{F}$ where the trial "basis" is given by the set of gridfunctions $(f_j(\mathbf{y}))_{j\in[J]}$ evaluated at the data and only the test function basis $(\phi_k)_{k\in[K]}$ is known analytically. In this way, inner products appearing in $\mathcal{R}(\mathbf{w};\phi)$ must be approximated numerically, implying that the accuracy of the recovered weights $\widehat{\mathbf{w}}$ is ultimately limited by the quadrature scheme used to discretize inner products. Using Lemma 2 below, we show that the correct coefficients $\mathbf{w}^\star$ may be recovered to effective machine precision accuracy (given by the tolerance of the forward ODE solver) from noise-free trajectories $\mathbf{y}$ by discretizing (2.2) using the trapezoidal rule and choosing ϕ to decay smoothly to zero at the boundaries of its support. Specifically, in this article we demonstrate this fact by choosing test functions from a particular family of unimodal piecewise polynomials $\mathcal{S}$ defined in (2.6).

Having chosen a quadrature scheme, the next accuracy barrier is presented by measurement noise, introducing randomness into the residuals $\mathcal{R}(\mathbf{w};\phi)$. Numerical integration then couples residuals $\mathcal{R}(\mathbf{w};\phi_1)$ and $\mathcal{R}(\mathbf{w};\phi_2)$ whenever ϕ_1 and ϕ_2 have overlapping support. In this way, $\mathcal{R}(\mathbf{w};\phi)$ does not have an ideal error structure for least squares but may be amenable to generalized least squares. Below we analyze the distribution of the residuals $\mathcal{R}(\mathbf{w};\phi)$ to arrive at a generalized least squares approach where an approximate covariance matrix can be computed directly from the test functions. This analysis also suggests that placing test functions near steep gradients in the dynamics may improve recovery; hence we develop a derivative-free method for adaptively clustering test functions near steep gradients.

REMARK 1. The weak formulation of the dynamics introduces a wealth of information: given M timepoints $\mathbf{t} = (t_m)_{m\in[M]}$, (2.2) affords $K = M(M-1)/2$ residuals over all possible supports $(a,b) \subset \mathbf{t}\times\mathbf{t}$ with $a < b$. Of course, one could also assimilate the responses of multiple families of test functions $\big(\{\phi^1_k\}_{k\in[K_1]}, \{\phi^2_k\}_{k\in[K_2]}, \dots\big)$; however, the computational complexity of such an exhaustive approach quickly becomes intractable. We stress that even with large noise, our proposed method identifies the correct nonlinearities with accurate weight recovery while keeping the number of test functions lower than the number of timepoints ($K < M$).

2.2. Algorithm: WSINDy. We state here the WSINDy algorithm in full generality. We propose a generalized least squares approach with approximate covariance matrix Σ. Below we derive a particular choice of Σ which utilizes the action of the test functions $(\phi_k)_{k\in[K]}$ on the data $\mathbf{y}$. Sequential thresholding on the weight coefficients $\mathbf{w}$ with thresholding parameter λ is used to enforce sparsity, where $\lambda \le \min_{\mathbf{w}^\star\neq 0}|\mathbf{w}^\star|$ is necessary for recovery. Lastly, an ℓ_2-regularization term with coefficient γ is included for problems involving rank deficiency. Methods of choosing optimal values of λ and γ directly from a given dataset do exist, for instance, by selecting the optimal position in a Pareto front [5]; however, this is not the focus of our current study, and thus we select values that work across multiple examples. Specifically, in the experiments below we set $\gamma = 0$ with the exception of the nonlinear pendulum and the five-dimensional linear system, examples which show that regularization can be used to discover dynamics from excessively large libraries. For noise-free data the

algorithm is only weakly dependent on λ, and so we use $\lambda = 0.001$, while for noisy data we set $\lambda = \frac{1}{4}\min_{\mathbf{w}^\star \neq 0} |\mathbf{w}^\star|$.

$\widehat{\mathbf{w}} = \mathbf{WSINDy}\left(\mathbf{y}, \mathbf{t}\ ;\ (\phi_k)_{k\in[K]},\ (f_j)_{j\in[J]},\ \Sigma, \lambda, \gamma\right)$:

1. Construct matrix of trial gridfunctions $\Theta(\mathbf{y}) = \left[f_1(\mathbf{y}) \,|\, \dots \,|\, f_J(\mathbf{y})\right]$.
2. Construct integration matrices $\mathbf{V}$, $\mathbf{V}'$ such that
$$\mathbf{V}_{km} = \Delta t \phi_k(t_m), \quad \mathbf{V}'_{km} = \Delta t \phi'_k(t_m).$$
3. Compute Gram matrix $\mathbf{G} = \mathbf{V}\Theta(\mathbf{y})$ and right-hand side $\mathbf{b} = -\mathbf{V}'\mathbf{y}$ so that $\mathbf{G}_{kj} = \langle \phi_k, f_j(\mathbf{y})\rangle$ and $\mathbf{b}_{kd} = -\langle \phi'_k, \mathbf{y}_d\rangle$.
4. Solve the generalized least squares problem with ℓ_2-regularization
$$\widehat{\mathbf{w}} = \text{argmin}_{\mathbf{w}} \left\{ (\mathbf{G}\mathbf{w} - \mathbf{b})^T \Sigma^{-1} (\mathbf{G}\mathbf{w} - \mathbf{b}) + \gamma^2 \|\mathbf{w}\|_2^2 \right\},$$
using sequential thresholding with parameter λ to enforce sparsity.

With this as our core algorithm, we can now consider a residual analysis (section 2.3) leading to a generalized least squares framework. We can also develop theoretical results related to the test functions (section 2.4), yielding a more thorough understanding of the impact of using uniform (section 2.4.1) and adaptive (section 2.4.2) placement of test functions along the time axis.

2.3. Residual analysis. Performance of WSINDy is determined by the behavior of the residuals
$$\mathcal{R}(\mathbf{w}; \phi_k) := (\mathbf{G}\mathbf{w} - \mathbf{b})_k \ \in\ \mathbb{R}^{1\times D},$$
denoted $\mathcal{R}(\mathbf{w}) \in \mathbb{R}^{K\times D}$ for the entire residual matrix. Here we analyze the residual for autonomous $\mathbf{F}$ to highlight key aspects for future analysis, as well as to arrive at an appropriate choice of approximate covariance Σ. We also provide a heuristic argument in favor of placing test functions near steep gradients in the dynamics.

A key difficulty in recovering the true weights $\mathbf{w}^\star$ is that for nonlinear systems the residual evaluated at the true weights $\mathbf{w}^\star$ is biased: $\mathbb{E}[\mathcal{R}(\mathbf{w}^\star)] \neq 0$. Any minimization of $\mathcal{R}$ thus introduces a bias in the recovered weights $\widehat{\mathbf{w}}$. Nevertheless, we can understand how different test functions impact the residual by linearizing around the true trajectory $\mathbf{x}(t)$ and isolating the dominant error terms:

$$\begin{aligned}
\mathcal{R}(\mathbf{w}; \phi_k) &= \langle \phi_k,\ \Theta(\mathbf{y})\mathbf{w}\rangle + \langle \phi'_k,\ \mathbf{y}\rangle \\
&= \langle \phi_k,\ \Theta(\mathbf{y})(\mathbf{w} - \mathbf{w}^\star)\rangle + \langle \phi_k,\ \Theta(\mathbf{y})\mathbf{w}^\star\rangle + \langle \phi'_k,\ \mathbf{y}\rangle \\
&= \langle \phi_k,\ \Theta(\mathbf{y})(\mathbf{w} - \mathbf{w}^\star)\rangle + \langle \phi_k,\ \mathbf{F}(\mathbf{y}) - \mathbf{F}(\mathbf{x})\rangle + \langle \phi'_k,\ \epsilon\rangle + I_k \\
&= \underbrace{\langle \phi_k,\ \Theta(\mathbf{y})(\mathbf{w} - \mathbf{w}^\star)\rangle}_{R_1} + \underbrace{\langle \phi_k,\ \epsilon\nabla\mathbf{F}(\mathbf{x})\rangle}_{R_2} + \underbrace{\langle \phi'_k,\ \epsilon\rangle}_{R_3} + I_k + \mathcal{O}(\epsilon^2),
\end{aligned}$$

where $\nabla F(\mathbf{x})_{dd'} = \frac{\partial \mathbf{F}_{d'}}{\partial \mathbf{x}_d}(\mathbf{x})$. The errors manifest in the following ways:

- R_1 is the misfit between $\mathbf{w}$ and $\mathbf{w}^\star$.
- R_2 results from measurement error in trial gridfunctions: $f_j(\mathbf{y}) = f_j(\mathbf{x}+\epsilon) \neq f_j(\mathbf{x})$.
- R_3 results from replacing $\mathbf{x}$ with $\mathbf{y} = \mathbf{x} + \epsilon$ in the left-hand side of (2.2).
- I_k is a deterministic integration error.
- $\mathcal{O}(\epsilon^2)$ is the remainder term in the truncated Taylor expansion of $\mathbf{F}(\mathbf{y})$ around $\mathbf{x}$:
$$\mathbf{F}(\mathbf{y}_m) = \mathbf{F}(\mathbf{x}(t_m)) + \epsilon_m \nabla\mathbf{F}(\mathbf{x}(t_m)) + \mathcal{O}(|\epsilon_m|^2).$$

Clearly, recovery of $\mathbf{F}$ when $\epsilon = 0$ is straightforward: R_1 and I_k are the only error terms; thus one only needs to select a quadrature scheme that ensures that the integration error I_k is negligible and $\widehat{\mathbf{w}} = \mathbf{w}^\star$ will be the minimizer. A primary focus of this study is the use of a specific family of piecewise polynomial test functions $\mathcal{S}$ defined below for which the trapezoidal rule is highly accurate (see Lemma 2). Figure 3.1 demonstrates this fact on noise-free data.

For $\epsilon > 0$, accurate recovery of $\mathbf{F}$ requires one to choose hyperparameters that exemplify the true misfit term R_1 by enforcing that the other error terms are of lower order. We look for $(\phi_k)_{k\in[K]}$ and $\Sigma = \mathbf{C}\mathbf{C}^T$ that approximately enforce $\mathbf{C}^{-1}\mathcal{R}(\mathbf{w}^\star) \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, justifying the least squares approach. In the next subsection we address the issue of approximating the covariance matrix, providing justification for using $\Sigma = \mathbf{V}'(\mathbf{V}')^T$. The following subsection provides a heuristic argument for how to reduce corruption from the error terms R_2 and R_3 by placing test functions near steep gradients in the data.

2.3.1. Approximate covariance Σ. Neglecting the deterministic integration error, which can be made small (see Lemma 2 below), and higher-order noise terms, the residual evaluated at the true weights is approximately

$$\mathcal{R}(\mathbf{w}^\star; \phi_k) \approx R_2 + R_3,$$

where $\mathbb{E}[R_2] = \mathbb{E}[R_3] = (0, \dots, 0)$ implies that $\mathbb{E}[\mathcal{R}(\mathbf{w}^\star)] = \mathbf{0}$ to leading order. Given the variances

$$\mathbb{V}[R_2] = \mathbb{V}[\langle \phi_k, \epsilon\nabla\mathbf{F}(\mathbf{x})\rangle] = \Delta t\, \sigma^2 \left(\left\| \phi_k \left|\nabla \mathbf{F}_1(\mathbf{x})\right| \right\|_2^2, \dots, \left\| \phi_k \left|\nabla \mathbf{F}_D(\mathbf{x})\right| \right\|_2^2 \right)$$

and

$$\mathbb{V}[R_3] = \mathbb{V}[\langle \phi_k', \epsilon\rangle] = \Delta t\, \sigma^2 \left(\|\phi_k'\|_2^2, \dots, \|\phi_k'\|_2^2 \right),$$

the true distribution of $\mathcal{R}(\mathbf{w}^\star)$ depends on $\mathbf{F}$, which is not known a priori. If it holds that $\|\phi_k'\|_2 \gg \|\phi_k |\nabla \mathbf{F}_d(\mathbf{x})|\|_2$, $d \in [D]$, a leading order approximation to $\mathrm{Cov}(\mathcal{R}(\mathbf{w}^\star))$ is

$$\Sigma := \mathbf{V}' \left(\mathbf{V}'\right)^T \propto \mathrm{Cov}(R_3),$$

using that $\mathrm{Cov}(R_3)_{ij} = \Delta t\sigma^2 \langle \phi_i', \phi_j' \rangle$. For this reason, we employ localized test functions and adopt the heuristic $\Sigma = \mathbf{V}'(\mathbf{V}')^T$ below.

2.3.2. Adaptive refinement. Next we show that by localizing ϕ_k around large $|\dot{\mathbf{x}}|$, we get an approximate cancellation of the error terms R_2 and R_3. Consider the one-dimensional case ($D = 1$) where m is an arbitrary time index and $\mathbf{y}_m = \mathbf{x}(t_m) + \epsilon$ is an observation. When $|\dot{\mathbf{x}}(t_m)|$ is large compared to ϵ, we approximately have

$$\mathbf{y}_m = \mathbf{x}(t_m) + \epsilon_m \approx \mathbf{x}(t_m + \delta t) \approx \mathbf{x}(t_m) + \delta t \mathbf{F}(\mathbf{x}(t_m)) \tag{2.4}$$

for some small δt, i.e., the perturbed value $\mathbf{y}_m$ lands close to the true trajectory $\mathbf{x}$ at the time $t_m + \delta t$. To understand the heuristic behind this approximation, let $t_m + \delta t$ be the point of intersection between the tangent line to $\mathbf{x}(t)$ at t_m and $\mathbf{x}(t_m) + \epsilon$. Then

$$\delta t = \frac{\epsilon}{\dot{\mathbf{x}}(t_m)};$$

hence $|\dot{\mathbf{x}}(t_m)| \gg \epsilon$ implies that $\mathbf{x}(t_m) + \epsilon$ will approximately lie on the true trajectory. As well, regions where $|\dot{\mathbf{x}}(t_m)|$ is small will not yield accurate recovery in the case of

noisy data, since perturbations are more likely to exit the relevant region of phase space. If we linearize $\mathbf{F}$ using the approximation (2.4) we get

$$\mathbf{F}(\mathbf{y}_m) \approx \mathbf{F}(\mathbf{x}(t_m)) + \delta t \mathbf{F}'(\mathbf{x}(t_m))\mathbf{F}(\mathbf{x}(t_m)) = \mathbf{F}(\mathbf{x}(t_m)) + \delta t \ddot{\mathbf{x}}(t_m). \tag{2.5}$$

Assuming ϕ_k is sufficiently localized around t_m, (2.4) also implies that

$$\langle \phi_k', \mathbf{x} \rangle + \underbrace{\langle \phi_k', \epsilon \rangle}_{R_3} = \langle \phi_k', \mathbf{y} \rangle \approx \langle \phi_k', \mathbf{x} \rangle + \delta t \langle \phi_k', \mathbf{F}(\mathbf{x}) \rangle ;$$

hence $R_3 \approx \delta t \langle \phi_k', \mathbf{F}(\mathbf{x}) \rangle$, while (2.5) implies

$$\begin{aligned} \langle \phi_k, \Theta(\mathbf{y})\mathbf{w} \rangle &= \underbrace{\langle \phi_k, \Theta(\mathbf{y})(\mathbf{w} - \mathbf{w}^\star) \rangle}_{=R_1} + \langle \phi_k, \mathbf{F}(\mathbf{y}) \rangle \\ &\approx \langle \phi_k, \Theta(\mathbf{y})(\mathbf{w} - \mathbf{w}^\star) \rangle + \langle \phi_k, \mathbf{F}(\mathbf{x}) \rangle + \underbrace{\delta t \langle \phi_k, \ddot{\mathbf{x}} \rangle}_{\approx R_2} \\ &= \langle \phi_k, \Theta(\mathbf{y})(\mathbf{w} - \mathbf{w}^\star) \rangle + \langle \phi_k, \mathbf{F}(\mathbf{x}) \rangle - \delta t \langle \phi_k', \mathbf{F}(\mathbf{x}) \rangle , \end{aligned}$$

having integrated by parts. Collecting the terms together yields that the residual takes the form

$$\mathcal{R}(\mathbf{w}; \phi_k) = \langle \phi_k', \mathbf{y} \rangle + \langle \phi_k, \Theta(\mathbf{y})\mathbf{w} \rangle \approx R_1,$$

and we see that R_2 and R_3 have effectively cancelled. In higher dimensions this interpretation does not appear to be as illuminating, but nevertheless, for any given coordinate $\mathbf{x}_d$, it does hold that terms in the error expansion vanish around points t_m where $|\dot{\mathbf{x}}_d|$ is large, precisely because $\mathbf{x}_d(t_m) + \epsilon \approx \mathbf{x}_d(t_m + \delta t)$.

2.4. Test function basis $(\phi_k)_{k\in[K]}$. Here we introduce a test function space $\mathcal{S}$ and quadrature scheme to minimize integration errors and enact the heuristic arguments above, which rely on ϕ_k having fast decay to its support boundaries and being sufficiently localized to ensure $\|\phi_k'\|_2^2 \gg \|\phi_k\|_2^2$. We define the space $\mathcal{S}$ of unimodal piecewise polynomials of the form

$$\phi(t) = \begin{cases} C(t-a)^p(b-t)^q & t \in [a,b], \\ 0 & \text{otherwise}, \end{cases} \tag{2.6}$$

where $(a,b) \subset \mathbf{t} \times \mathbf{t}$ satisfies $a < b$ and $p, q \geq 1$. The normalization

$$C = \frac{1}{p^p q^q} \left(\frac{p+q}{b-a} \right)^{p+q}$$

ensures that $\|\phi\|_\infty = 1$. Functions $\phi \in \mathcal{S}$ are nonnegative, unimodal, and compactly supported in $[0,T]$ with $\lfloor \min\{p,q\} \rfloor - 1$ continuous derivatives. Larger p and q imply faster decay towards the endpoints of the support. For $p = q$, we refer to p as the degree of ϕ.

To ensure the integration error in approximating inner products $\langle f_j, \phi_k \rangle$ is negligible, we rely on the following lemma, which provides a bound on the error in discretizing the weak derivative relation

$$-\int \phi' f \, dt = \int \phi f' \, dt \tag{2.7}$$

using the trapezoidal rule for compactly supported ϕ. Following the lemma we introduce two strategies for choosing the parameters of the test functions $(\phi_k)_{k\in[K]} \subset \mathcal{S}$.

LEMMA 2 (numerical error in weak derivatives). *Let f, ϕ have continuous derivatives of order p, and define $t_j = a + j\frac{b-a}{N} = a + j\Delta t$. If ϕ has roots $\phi(a) = \phi(b) = 0$ of multiplicity p, then*

$$\frac{\Delta t}{2}\sum_{j=0}^{N-1}\left[g(t_j) + g(t_{j+1})\right] = \mathcal{O}(\Delta t^{p+1}), \tag{2.8}$$

where $g(t) = \phi'(t)f(t) + \phi(t)f'(t)$. In other words, the composite trapezoidal rule discretizes the weak derivative relation (2.7) to order $p+1$.

Proof. This is a simple consequence of the Euler-Maclaurin formula. If $g : [a, b] \to \mathbb{C}$ is a smooth function, then the following asymptotic expansion holds:

$$\frac{\Delta t}{2}\sum_{j=0}^{N-1}\left[g(t_j) + g(t_{j+1})\right] \sim \int_a^b g(t)\,dt + \sum_{k=1}^{\infty}\frac{\Delta t^{2k}B_{2k}}{(2k)!}\left(g^{(2k-1)}(b) - g^{(2k-1)}(a)\right),$$

where B_{2k} are the Bernoulli numbers. The asymptotic expansion provides corrections to the trapezoidal rule that realize machine precision accuracy up until a certain value of k, after which terms in the expansion grow and the series diverges [6, Chapter 3]. In our case, $g(t) = \phi'(t)f(t) + \phi(t)f'(t)$, where the root conditions on ϕ imply that

$$\int_a^b g(t)\,dt = 0 \quad \text{and} \quad g^{(k)}(b) = g^{(k)}(a) = 0, \quad 0 \le k \le p-1.$$

So for p odd, we have that

$$\begin{aligned}\frac{\Delta t}{2}\sum_{j=0}^{N-1}\left[g(t_j) + g(t_{j+1})\right] &\sim \sum_{k=(p+1)/2}^{\infty}\frac{\Delta t^{2k}B_{2k}}{(2k)!}\left(g^{(2k-1)}(b) - g^{(2k-1)}(a)\right)\\ &= \frac{B_{p+1}}{(p+1)!}(\phi^{(p)}(b)f(b) - \phi^{(p)}(a)f(a))\Delta t^{p+1} + \mathcal{O}\left(\Delta t^{p+2}\right).\end{aligned}$$

For even p, the leading term is $\mathcal{O}(\Delta t^{p+2})$ with a slightly different coefficient. ∎

For $\phi \in \mathcal{S}$ with $p = q$, the exact leading order error in term in (2.8) is

$$\frac{2^p\,B_{p+1}}{p+1}\big(f(b) - f(a)\big)\Delta t^{p+1}, \tag{2.9}$$

which is negligible for a wide range of reasonable p and Δt values. The Bernoulli numbers eventually start growing like p^p, but for smaller values of p they are moderate. For instance, with $\Delta t = 0.1$ and $f(b) - f(a) = 1$, this error term is $o(1)$ up until $p = 85$, where it takes the value 0.495352, while for $\Delta t = 0.01$, the error is below machine precision for all p between 7 and 819. For these reasons, in what follows we choose test functions $(\phi_k)_{k\in[K]} \subset \mathcal{S}$ and discretize all integrals using the trapezoidal rule. Unless otherwise stated, each function ϕ_k satisfies $p = q$ and so is fully determined by the tuple $\{p_k, a_k, b_k\}$ indicating its polynomial degree and support. In the next two subsections we propose two different strategies for determining ϕ_k using the data $\mathbf{y}$.

2.4.1. Strategy 1: Uniform grid. The simplest strategy for choosing a basis of test functions $(\phi_k)_{k\in[K]} \subset \mathcal{S}$ is to place ϕ_k uniformly on the interval $[0, T]$ with fixed degree p and fixed support size

$$L := \#\{\mathbf{t} \cap \mathrm{supp}(\phi_k)\}$$

(i.e., L is the number of timepoints in $\mathbf{t}$ that ϕ_k is supported on). The triple (L, p, K) then defines the scheme, where each piece effects the distribution of the residual $\mathcal{R}(\mathbf{w})$.

Step 1: *Choosing* L. Heuristically, the support size of ϕ_k relates to the Fourier transform of the data. If $\text{supp}(\phi_k)$ is small compared to the dominant wavemodes in the dynamics, then high-frequency noise will dominate the values of the inner products $\langle \phi_k', \mathbf{y} \rangle$. If $\text{supp}(\phi_k)$ is much larger than the dominant wavemodes, then too much averaging may occur, leading to unresolved dynamics. A natural choice is then to set L equal to the period of a known active wavemode[1] k:

$$L = \left\lfloor \frac{1}{\Delta t} \frac{2\pi}{(2\pi T/k)} \right\rfloor = \left\lfloor \frac{M}{k} \right\rfloor.$$

In the noise-free and small-noise experiments below we set $L = \lfloor \frac{M}{25} \rfloor$ and leave optimal selection of L based on Fourier analysis to future work.

Step 2: *Determining* p. In light of the derivation above of the approximate covariance matrix $\Sigma = \mathbf{V}'(\mathbf{V}')^T$, we define the parameter $\rho := \|\phi_k'\|_2 / \|\phi_k\|_2$, which serves as an estimate for the ratio $\sqrt{\mathbb{V}[R_3]/\mathbb{V}[R_2]}$ between the standard deviations of the two dominant error terms R_3 and R_2 in the residual $\mathcal{R}(\mathbf{w}^\star)$. Larger ρ indicates better agreement with the approximate covariance matrix Σ, since $\Sigma \propto \text{Cov}(R_3)$. Furthermore, for $\phi_k \in \mathcal{S}$ we have the exact formula

$$\rho^2 = \frac{8p^2}{(b-a)^2} \left(\frac{\Gamma(2p-1)\Gamma(2p+\frac{1}{2})}{\Gamma(2p+1)\Gamma(2p+\frac{3}{2})} \right) = \frac{p}{(b-a)^2} \left(\frac{4p+1}{p-\frac{1}{2}} \right),$$

where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$ is the gamma function. Given $\rho^2 \geq (5 + 2\sqrt{6})/(b-a)^2$, a polynomial degree p may be selected from ρ using the formula

$$p = \left\lfloor \frac{1}{8} \left(((b-a)^2\rho^2 - 1) + \sqrt{((b-a)^2\rho^2 - 1)^2 - 8(b-a)^2\rho^2} \right) \right\rfloor.$$

Step 3: *Determining* K. Next we introduce the shift parameter $s \in [0, 1]$ defined by

$$s := \phi_k(t^*) \;\text{ s.t. }\; \phi_k(t^*) = \phi_{k+1}(t^*),$$

which determines K from p and L. In words, s is the height of intersection between ϕ_k and ϕ_{k+1} and measures the amount of overlap between successive test functions. More overlap increases the correlation between rows in the residual $\mathcal{R}(\mathbf{w})$ and hence leads to larger off-diagonal elements in the covariance matrix Σ. Larger s implies that neighboring functions overlap on more points, with $s = 1$ indicating that $\phi_k = \phi_{k+1}$. Specifically, neighboring test functions overlap on $\lfloor L(1 - \sqrt{1 - s^{1/p}}) \rfloor$ timepoints. In Figures 3.2 and 3.3 we vary the parameters ρ and s and observe that results agree with intuition: larger ρ (better agreement with Σ) and larger s (more test functions) lead to better recovery of $\mathbf{w}^\star$. We summarize the uniform grid algorithm below.

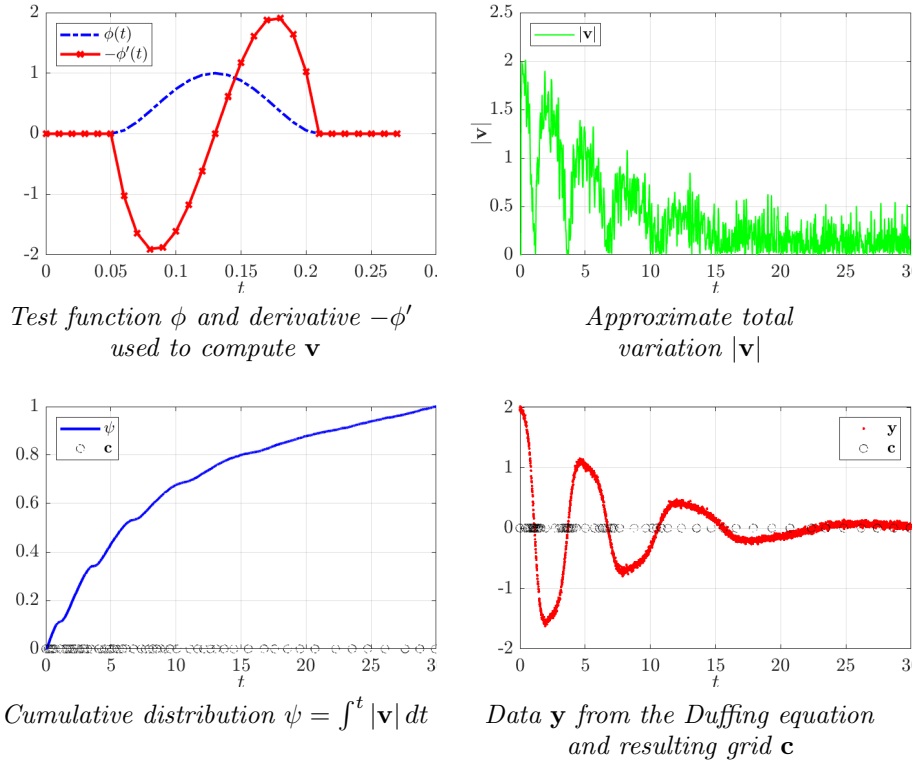
$\widehat{\mathbf{w}} = \textbf{WSINDy\_UG}\left(\mathbf{y}, \mathbf{t}\ ;\ (f_j)_{j\in[J]}, L, \rho, s, \lambda, \gamma\right)$:

1. Construct matrix of trial gridfunctions $\Theta(\mathbf{y}) = \left[f_1(\mathbf{y}) \,|\, \dots \,|\, f_J(\mathbf{y}) \right]$.
2. Construct integration matrices $\mathbf{V}$, $\mathbf{V}'$ such that

$$\mathbf{V}_{km} = \Delta t \phi_k(t_m), \quad \mathbf{V}'_{km} = \Delta t \phi_k'(t_m)$$

with the test functions $(\phi_k)_{k\in[K]}$ determined by L, ρ, s as described above.

[1] Such that $\mathcal{F}_k(\mathbf{y}) := \sum_{j=0}^{M-1} \mathbf{y}_m e^{-2\pi i j k/M}$ is not negligible.



Test function ϕ and derivative $-\phi'$ used to compute $\mathbf{v}$

Approximate total variation $|\mathbf{v}|$

Cumulative distribution $\psi = \int^t |\mathbf{v}|\, dt$

Data $\mathbf{y}$ from the Duffing equation and resulting grid $\mathbf{c}$

FIG. 2.1. *Adaptive grid construction used on data from the Duffing equation with* 10% *noise* ($\sigma_{NR} = 0.1$). *As desired, the centers* $\mathbf{c}$ *are clustered near steep gradients in the dynamics despite large measurement noise. (Note* $-\phi(t)'/10$ *is plotted in the upper-left instead of* $-\phi(t)'$ *in order to visualize both* ϕ *and* ϕ'.*)*

3. Compute Gram matrix $\mathbf{G} = \mathbf{V}\Theta(\mathbf{y})$ and right-hand side $\mathbf{b} = -\mathbf{V}'\mathbf{y}$ so that $\mathbf{G}_{kj} = \langle \phi_k, f_j(\mathbf{y}) \rangle$ and $\mathbf{b}_{kd} = -\langle \phi_k', \mathbf{y}_d \rangle$.
4. Compute approximate covariance and Cholesky factorization $\Sigma = \mathbf{V}'(\mathbf{V}')^T = \mathbf{C}\mathbf{C}^T$.
5. Solve the generalized least squares problem with ℓ_2-regularization

$$\widehat{\mathbf{w}} = \text{argmin}_{\mathbf{w}} \left\{ (\mathbf{G}\mathbf{w} - \mathbf{b})^T \Sigma^{-1} (\mathbf{G}\mathbf{w} - \mathbf{b}) + \gamma^2 \|\mathbf{w}\|_2^2 \right\},$$

using sequential thresholding with parameter λ to enforce sparsity.

2.4.2. Strategy 2: Adaptive grid. Motivated by the arguments above, we now introduce an algorithm for constructing a test function basis localized near points of large change in the dynamics. This occurs in three steps: (1) construct a weak approximation to the derivative of the dynamics $\mathbf{v} \approx \dot{\mathbf{x}}$, (2) sample K points $\mathbf{c}$ from a cumulative distribution ψ with density proportional to the total variation $|\mathbf{v}|$, and (3) construct test functions centered at $\mathbf{c}$ using a width-at-half-max parameter r_{whm} to determine the parameters (p_k, a_k, b_k) of each function ϕ_k. Each of these steps is numerically stable and carried out independently along each coordinate of the dynamics. A visual diagram is provided in Figure 2.1.

Step 1: *Weak derivative approximation.* Define $\mathbf{v} := -\mathbf{V}'_w \mathbf{y}$, where the matrix $-\mathbf{V}'_w$ enacts a linear convolution with the derivative of a chosen test function $\phi \in \mathcal{S}$ of degree p_w and support size L_w so that

$$\mathbf{v}_m = -\langle \phi', \mathbf{y} \rangle = \langle \phi, \dot{\mathbf{y}} \rangle \approx \dot{\mathbf{y}}_m.$$

The parameters L_w and p_w are chosen by the user, with $L_w = 5$ and $p_w \geq 2$ corresponding to taking a centered finite difference derivative with a 3-point stencil. Smaller p_w results in more smoothing and minimizes the corruption from noise while still accurately locating steep gradients in the dynamics. For the examples below we arbitrarily[2] use $p_w = 2$ and $L_w = 17$.

Step 2: *Selecting* $\boldsymbol{c}$. Having computed $\mathbf{v}$, define ψ to be the cumulative sum of $|\mathbf{v}|$ normalized so that $\max \psi = 1$. In this way ψ is a valid cumulative distribution function with density proportional to the total variation of $\mathbf{y}$. We then find $\mathbf{c}$ by sampling from ψ. Let $U = [0, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}]$ with K being the number of the test functions; we then define $\mathbf{c} = \psi^{-1}(U)$, or numerically,

$$c_k = \min\{t \in \mathbf{t} \ : \ \psi(t) \geq U_k\}.$$

This stage requires the user to select the number of test functions K.

Step 3: *Construction of test functions* $(\phi_k)_{k \in [K]}$. Having chosen the location c_k of the centerpoint for each test function ϕ_k, we are left to choose the degree p_k of the polynomial and the supports $[a_k, b_k]$. The degree is chosen according to the width-at-half-max parameter r_{whm}, which specifies the difference in timepoints between each center c_k and $\arg_t\{\phi_k(t) = 1/2\}$, while the supports are chosen such that $\phi_k(b_k - \Delta t) = 10^{-16}$. This gives us a nonlinear system of two equations in two unknowns which can be easily solved (i.e., using `fzero` in MATLAB). This can be done for one reference test functions and the rest of the weights obtained by translation. The optimal value of r_{whm} depends on the timescales of the dynamics and can be chosen from the data using the Fourier transform as in the uniform grid case; however, for simplicity we set $r_{whm} = \lfloor M/100 \rfloor$ in the large-noise examples below.

The adaptive grid WSINDy algorithm is summarized as follows:

$\widehat{\mathbf{w}} = \mathbf{WSINDy\_AG}\left(\mathbf{y}, \mathbf{t} \ ; \ (f_j)_{j \in [J]}, p_w, L_w, K, r_{whm}, \lambda, \gamma\right)$:

1. Construct matrix of trial gridfunctions $\Theta(\mathbf{y}) = \left[f_1(\mathbf{y}) \,|\, \ldots \,|\, f_J(\mathbf{y})\right]$.
2. Construct integration matrices $\mathbf{V}$, $\mathbf{V}'$ such that
$$\mathbf{V}_{km} = \Delta t \phi_k(t_m), \quad \mathbf{V}'_{km} = \Delta t \phi'_k(t_m),$$
with test functions $(\phi_k)_{k \in [K]}$ determined by p_w, L_w, K, r_{whm} as described above.
3. Compute Gram matrix $\mathbf{G} = \mathbf{V}\Theta(\mathbf{y})$ and right-hand side $\mathbf{b} = -\mathbf{V}'\mathbf{y}$ so that $\mathbf{G}_{kj} = \langle \phi_k, f_j(\mathbf{y}) \rangle$ and $\mathbf{b}_{kd} = -\langle \phi'_k, \mathbf{y}_d \rangle$.
4. Compute approximate covariance and Cholesky factorization $\Sigma = \mathbf{V}'(\mathbf{V}')^T = \mathbf{C}\mathbf{C}^T$.
5. Solve the generalized least squares problem with ℓ_2-regularization
$$\widehat{\mathbf{w}} = \text{argmin}_{\mathbf{w}} \left\{ (\mathbf{G}\mathbf{w} - \mathbf{b})^T \Sigma^{-1} (\mathbf{G}\mathbf{w} - \mathbf{b}) + \gamma^2 \|\mathbf{w}\|_2^2 \right\},$$
using sequential thresholding with parameter λ to enforce sparsity.

[2] We find that a lower-degree test function with small support effectively locates steep gradients in noisy trajectories.

3. Numerical experiments. We now show that WSINDy is capable of recovering the correct dynamics to high accuracy over a range of noise levels. We examine the systems in Table 1 which exhibit several canonical dynamics, namely growth and decay, nonlinear oscillations and chaotic dynamics, in dimensions $D \in \{2, 3, 5\}$. To generate true trajectory data we use `ode45` in MATLAB with absolute and relative tolerance 10^{-10} and collect M samples uniformly[3] in time with sampling rate Δt. The parameters M and Δt are chosen to provide a balance between illustrating ODE behaviors and avoiding an overabundance of observations. Gaussian white noise with mean zero and variance σ^2 is added to the exact trajectories, where σ is computed by specifying a noise ratio σ_{NR} and setting

$$\sigma = \sigma_{NR} \frac{\|\mathbf{x}\|_F}{\sqrt{MD}}, \tag{3.1}$$

where the Frobenius norm of a matrix $\mathbf{x} \in \mathbb{R}^{M \times D}$ is defined by

$$\|\mathbf{x}\|_F := \sqrt{\sum_{m=1}^{M} \sum_{d=1}^{D} |\mathbf{x}_{md}|^2}.$$

The ratio of noise to signal is then approximately equal to the square root of the variance: $\|\epsilon\|_F / \|\mathbf{x}\|_F \approx \sigma$.

We measure the accuracy in the recovered dynamical system using the relative $\|\cdot\|_F$ error in the recovered coefficients,

$$E_2(\widehat{\mathbf{w}}) = \frac{\|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_F}{\|\mathbf{w}^\star\|_F}, \tag{3.2}$$

and the relative $\|\cdot\|_F$ error between the noise-free data $\mathbf{x}$ and the data-driven dynamics $\mathbf{x}_{dd}$ along the same timepoints:

$$\mathcal{E}_2(\mathbf{x}_{dd}) = \frac{\|\mathbf{x}_{dd} - \mathbf{x}\|_F}{\|\mathbf{x}\|_F}. \tag{3.3}$$

The collection of ODEs in Table 1 are all first-order autonomous systems; however, they exhibit a diverse range of dynamics. The Linear 5D system (for $\beta < 0$) and Duffing's equation are both examples of damped oscillators, showing that WSINDy is able to discern whether such motion is governed by linear or nonlinear coupling between variables. For $\beta > 0$, the Linear 5D system exhibits exponential growth. The van der Pol oscillator, Lotka–Volterra system, and nonlinear pendulum demonstrate that a stable limit cycle with abrupt changes may manifest from vastly different nonlinear mechanisms, which turn out to be identifiable using the weak form. Finally, the Lorenz system exhibits deterministic chaos, and hence the dynamics cover a wide range of Fourier modes, which easily become corrupted with noise.

3.1. Noise-free data. The goal of the following noise-free experiments is to demonstrate convergence of the recovered weights $\widehat{\mathbf{w}}$ to the true weights $\mathbf{w}^\star$ to within the accuracy tolerance of the ODE solver (fixed 10^{-10} throughout). In light of Lemma 2, this should occur as the decay rate of the test functions $(\phi_k)_{k \in [K]}$ is increased, which for test functions in class $\mathcal{S}$ (see (2.6)) is realized by increasing the polynomial degree

[3] We leave a detailed study of nonuniform time sampling to future work.

TABLE 1
ODEs used in numerical experiments. For Linear 5D, Duffing, van der Pol, and Lotka–Volterra we measure the accuracy in the recovered system as the parameter β varies (see Table 2).

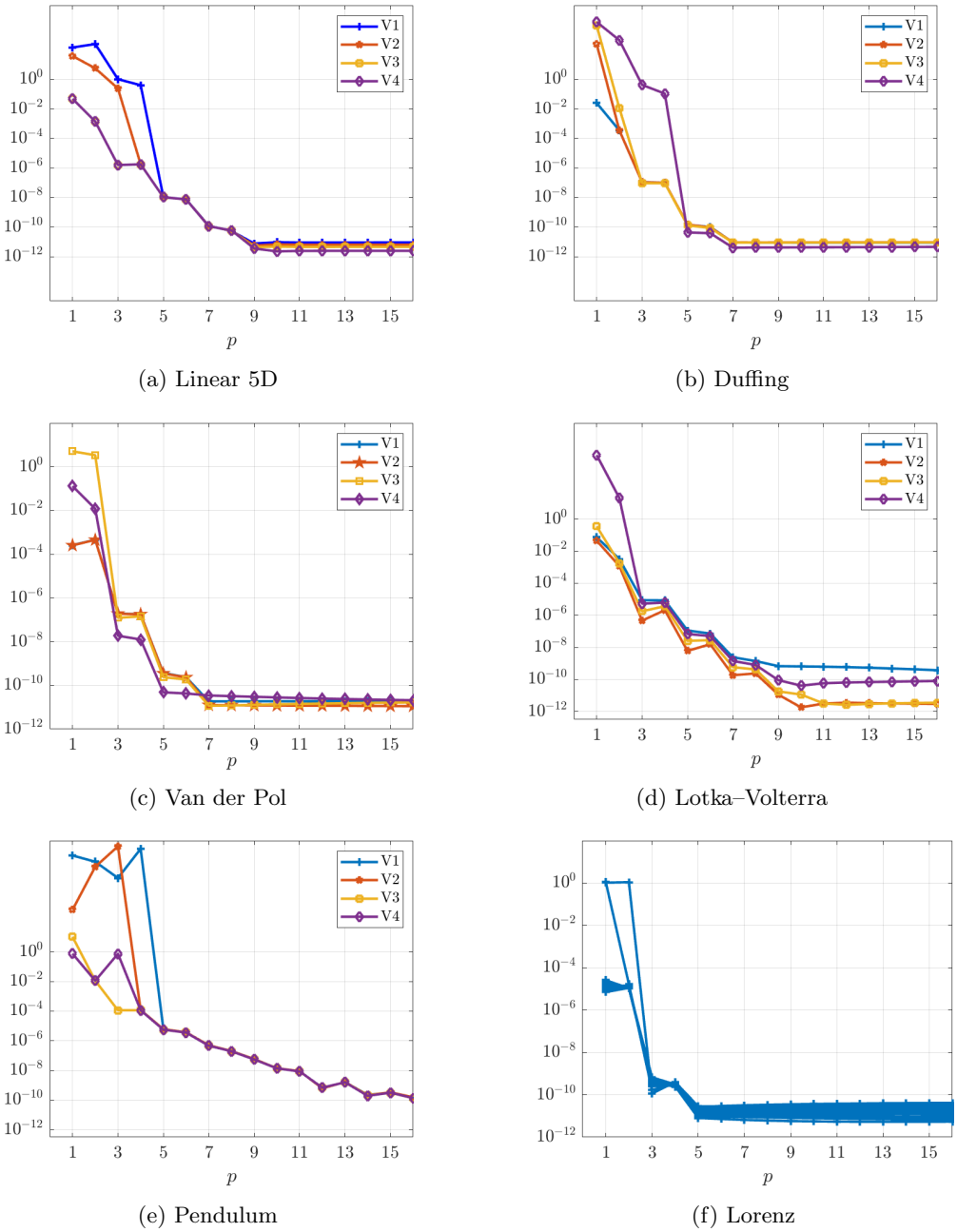
Name	Governing equations	M	Δt
Linear 5D	$\begin{cases}\dot{x}_1 = -x_5 + \beta x_1 + x_2,\\ \dot{x}_i = -x_{i-1} + \beta x_i + x_{i+1}, \quad i = 2,3,4\\ \dot{x}_5 = -x_4 + \beta x_5 + x_1\end{cases}$	1401	0.025
Duffing	$\begin{cases}\dot{x}_1 = x_2,\\ \dot{x}_2 = -0.2x_2 - 0.2x_1 - \beta x_1^3\end{cases}$	3001	0.01
Van der Pol	$\begin{cases}\dot{x}_1 = x_2,\\ \dot{x}_2 = \beta x_2(1 - x_1^2) - x_1\end{cases}$	3001	0.01
Lotka–Volterra	$\begin{cases}\dot{x}_1 = 3x_1 - \beta x_1 x_2,\\ \dot{x}_2 = \beta x_1 x_2 - 6x_2\end{cases}$	1001	0.01
Nonlinear pendulum	$\begin{cases}\dot{x}_1 = x_2,\\ \dot{x}_2 = -\sin(x_1)\end{cases}$	501	0.1
Lorenz	$\begin{cases}\dot{x}_1 = 10(x_2 - x_1),\\ \dot{x}_2 = x_1(28 - x_3) - x_2,\\ \dot{x}_3 = x_1 x_2 - \frac{8}{3}x_3\end{cases}$	10001	0.001

TABLE 2
Specifications for parameters used in illustrating simulations in Figure 3.1.

ODE	β	$\mathbf{x}(0)$	L	ΔL	$J(=K)$
Linear 5D	$(-0.3, -0.2, -0.1, 0.1)$	$(10,0,0,0,0)^T$	57	5	252
Duffing	$(0.01, 0.1, 1, 10)$	$(0,2)^T$	121	99	29
Van der Pol	$(0.01, 0.1, 1, 10)$	$(0,1)^T$	121	99	29
Lotka–Volterra	$(0.005, 0.01, 0.1, 1)$	$(1,1)^T$	41	33	29
Pendulum	—	$x_2(0) = 0$, $x_1(0) \in \{\frac{15}{16}\pi, \frac{10}{16}\pi, \frac{5}{16}\pi, \frac{1}{16}\pi\}$	21	16	29
Lorenz	—	$\sim U_{[-15,15]^2 \times [10,40]}$	401	141	68

p. Hence, over the range of parameter values in Table 2, for each system we test convergence as p increases. We use the uniform grid approach with shift parameter s chosen such that the number of test functions equals to the number of trial functions ($K = J$), resulting in square Gram matrices $\mathbf{G} = \mathbf{V}\Theta(\mathbf{y})$. The support of the basis functions along the timegrid $\mathbf{t}$ is set to $L = \lfloor \frac{M}{25} \rfloor$ points. The data-driven trial basis $(f_j)_{j\in[J]}$ includes all monomials in the state variables up to degree 5 as well as the trigonometric terms $\cos(n\mathbf{y}_d)$, $\sin(n\mathbf{y}_d)$ for $n = 1, 2$ and $d \in [D]$. We set the regularization parameter to zero ($\gamma = 0$), with the exception of the nonlinear pendulum, where $\gamma = 10^{-8}$, and the sparsity threshold to $\lambda = 0.001$. We note that a nonzero γ is always necessary to discover the nonlinear pendulum from combined trigonometric and polynomial libraries since $\sin(x_1)$ is well-approximated by polynomial terms; however, the same is not true for low-order polynomial systems. In cases considered here, sequential thresholding successfully removes trigonometric library terms for ODE systems with polynomial dynamics despite initially ill-conditioned Gram matrices $\mathbf{G}$ resulting from combining polynomial and trigonometric terms.

Figure 3.1 shows that in the limit of large p, WSINDy recovers the correct weight matrix $\mathbf{w}^\star$ of each system in Table 1 to an accuracy of $\mathcal{O}(10^{-10})$. For the Linear 5D system, we vary the growth/decay parameter, showing that the system is identifiable to high accuracy despite an excessively large trial library (252 terms). For Duffing's



(a) Linear 5D (b) Duffing

(c) Van der Pol (d) Lotka–Volterra

(e) Pendulum (f) Lorenz

FIG. 3.1. *Noise-free data* ($\sigma_{NR} = 0$)*: plots of relative coefficient error* $E_2(\widehat{\mathbf{w}})$ *(defined in* (3.2)*) vs. p. V1-V4 indicate different ODE parameters (see Table* 2*). For the Lorenz system the parameters are fixed, and* 40 *different initial conditions are sampled from a uniform distribution. In each case, the recovered coefficients* $\widehat{\mathbf{w}}$ *rapidly converge to within the accuracy of the ODE solver* (10^{-10}).

equation and the van der Pol oscillator, the same convergence trend is observed for β values spanning several orders of magnitude. Accuracy is slightly worse for the Lotka–Volterra equation when $\beta = 0.005$, which corresponds to highly infrequent predator-

prey interactions and leads to solutions with large amplitudes and gradients. For the nonlinear pendulum, we test that WSINDy is able to identify the $\sin(x_1)$ nonlinearity for both large and small initial amplitudes, noting that $x_1(0) = \frac{15}{16}\pi \approx \pi$ produces strongly nonlinear oscillations, while $x_1(0) = \frac{1}{16}\pi$ produces small-angle oscillations where $\sin(x_1) \approx x_1$. In addition, for the pendulum we use fewer samples ($M = 501$) and a larger time step $\Delta t = 0.1$ and hence observe a decreased convergence rate. For the Lorenz equations we vary the initial conditions, generating 40 random initial conditions from a region covering the strange attractor, and show convergence over all cases.

3.2. Small-noise regime. We now turn to the case of low to moderate noise levels, examining a noise ratio σ_{NR} in the range $[10^{-5}, 0.04]$ for the van der Pol oscillator and Duffing's equation. We examine $\rho \in [1, 7]$ and $s \in [0.3, 0.95]$, where $\rho := \|\phi_k'\|_2 / \|\phi_k\|_2$ and s is the height of intersection of two neighboring test functions ϕ_k and ϕ_{k+1} (with $s = 1$ leading to $\phi_k = \phi_{k+1}$ and $s = 0$ indicating $\text{supp}(\phi_k) \cap \text{supp}(\phi_{k+1}) = \emptyset$). Using the analysis from section 2.3, increasing ρ affects the distribution of the residual $\mathcal{R}(\mathbf{w})$ by magnifying the portion $R_3 = \langle \phi_k', \epsilon \rangle$ that is linear in the noise. For $\phi \in \mathcal{S}$, larger ρ corresponds to a higher polynomial degree p, with $\rho \in [1, 7]$ leading to $p \in [2, 98]$. Larger shift parameter s corresponds to more test functions (higher K) but also to higher correlation between rows in $\mathbf{G}$, as $\langle \phi_k, f_j(\mathbf{y}) \rangle \approx \langle \phi_{k+1}, f_j(\mathbf{y}) \rangle$ when the supports of ϕ_k and ϕ_{k+1} sufficiently overlap. Here $s \in [0.3, 0.95]$ corresponds to $K \in [14, 451]$. We again use the uniform grid approach with $\gamma = 0$ and $\lambda = \frac{1}{4} \min_{\mathbf{w}_j^\star \neq 0} |\mathbf{w}_j^\star|$. For each system we generate 200 instantiations of noise and record the coefficient error over the range of s and ρ values.

From Figures 3.2 and 3.3 we observe two properties. Firstly, the coefficient error $E_2(\widehat{\mathbf{w}})$ monotonically deceases with increasing s and ρ; hence accurate recovery re-

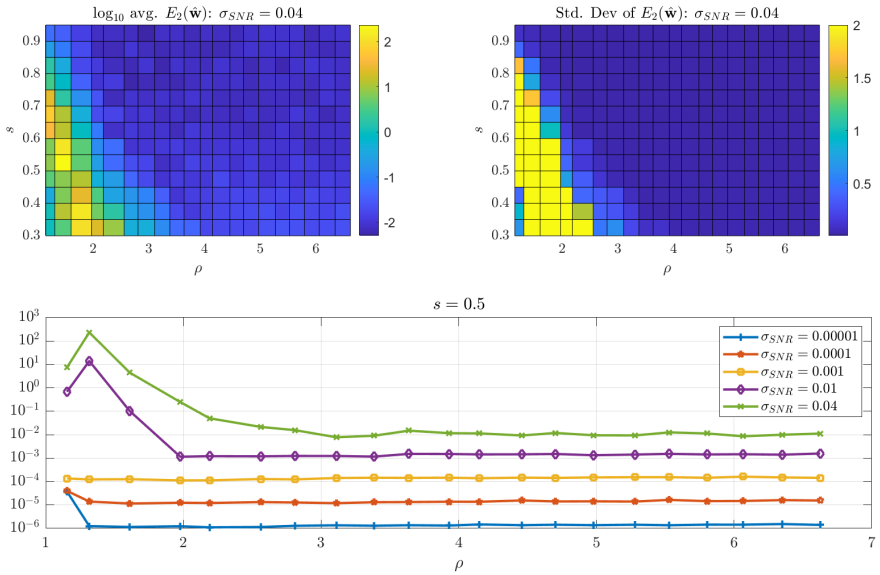


FIG. 3.2. *Small-noise regime: dynamic recovery of the Duffing equation with $\beta = 1$. Top: heat map of the $\log_{10}$ average error $E_2(\widehat{\mathbf{w}})$ (left) and sample standard deviation of $E_2(\widehat{\mathbf{w}})$ (right) over 200 instantiations of noise with $\sigma_{NR} = 0.04$ (4% noise) vs. ρ and s. Bottom: $E_2(\widehat{\mathbf{w}})$ vs. ρ for fixed $s = 0.5$ and various σ_{NR}. For $\rho > 3$ the average error is roughly an order of magnitude below σ_{NR}.*

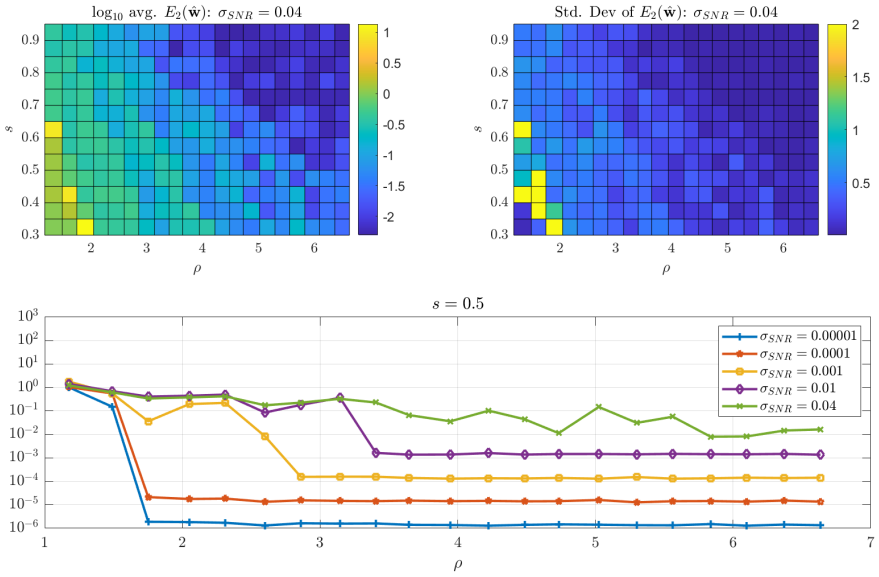


FIG. 3.3. *Small-noise regime: dynamic recovery of the van der Pol oscillator with* $\beta = 4$. *Top: heat map of the* $\log_{10}$ *average error* $E_2(\widehat{\mathbf{w}})$ *(left) and sample standard deviation of* $E_2(\widehat{\mathbf{w}})$ *(right) over* 200 *instantiations of noise with* $\sigma_{NR} = 0.04$ *(4% noise) vs.* ρ *and* s. *Bottom:* $E_2(\widehat{\mathbf{w}})$ *vs.* ρ *for fixed* $s = 0.5$ *and various* σ_{NR}. *Similar to the Duffing equation, average error falls to roughly an order of magnitude below* σ_{NR}, *although for van der Pol this regime is reached when* $\rho \approx 6$.

quires sufficient overlap between test functions (large enough shift parameter s) and sufficiently localized test functions that amplify the portion of the residual that is linear in the noise. Secondly, for large enough ρ and s, the error in the coefficients scales linearly with σ_{NR}, leading to an accuracy of $E_2(\widehat{\mathbf{w}}) \approx 0.1\sigma_{NR}$, or $-\log_{10}(0.1\sigma_{NR})$ significant digits in the recovered coefficients. In Appendix A we show that this second property does not hold for standard SINDy; in particular, the method of differentiation must change depending on the noise level in order to reach a desired accuracy.

3.3. Large-noise regime. Figures 3.4 to 3.9 show that adaptive placement of test functions (Strategy 2) can be employed to discover dynamics in the large-noise regime with fewer test functions. We test that each system in Table 1 can be discovered under $\sigma_{NR} = 0.1$ (10% noise) from only 250 test functions distributed near steep gradients in $\mathbf{y}$, which are located using the scheme in section 2.4.2 with $p_w = 2$ and $L_w = 17$. We set the width-at-half-max of the test functions to $r_{whm} = \lfloor M/100 \rfloor$ timepoints. To exemplify the separation of scales and the severity of the corruption from noise, the noisy data $\mathbf{y}$, true data $\mathbf{x}$, and trajectories $\mathbf{x}_{dd}$ from the learned dynamical systems are shown in dynamo view and in phase space (for $D \leq 3$). We extend $\mathbf{x}_{dd}$ by 50% to show that the data-driven system captures the true limiting behavior. We set the sparsity to $\lambda = \frac{1}{4}\min_{\mathbf{w}^\star \neq 0} |\mathbf{w}^\star|$ and $\gamma = 0$ except in the Linear 5D and nonlinear pendulum examples, where $\gamma = \sqrt{\sigma_{NR}} \approx 0.32$. For the trial basis we use all monomials up to degree 5 in the state variables, and for the pendulum we include the trigonometric terms $\sin(k\mathbf{y}_d), \cos(k\mathbf{y}_d)$ for $k = 1, 2$ and $d = 1, 2$.

In each case the correct terms are identified with coefficient error $E_2(\widehat{\mathbf{w}}) < 10^{-2}$, in agreement with the trend $E_2(\widehat{\mathbf{w}}) \approx 0.1\sigma_{NR}$ observed in the small-noise regime. For the Linear 5D, Duffing, and Lotka–Volterra systems (Figures 3.4, 3.5, and 3.7) the data-driven trajectory $\mathbf{x}_{dd}$ is indistinguishable from the true data to the eye,

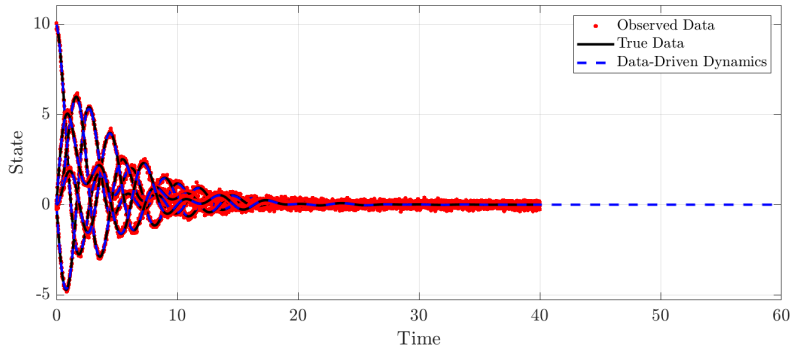


FIG. 3.4. *Large-noise regime: Linear 5D system with damping $\beta = -0.2$. All correct terms were identified with an error in the weights of $E_2(\widehat{\mathbf{w}}) = 0.0064$ and a trajectory error of $\mathcal{E}_2(\widehat{\mathbf{w}}) = 0.013$.*

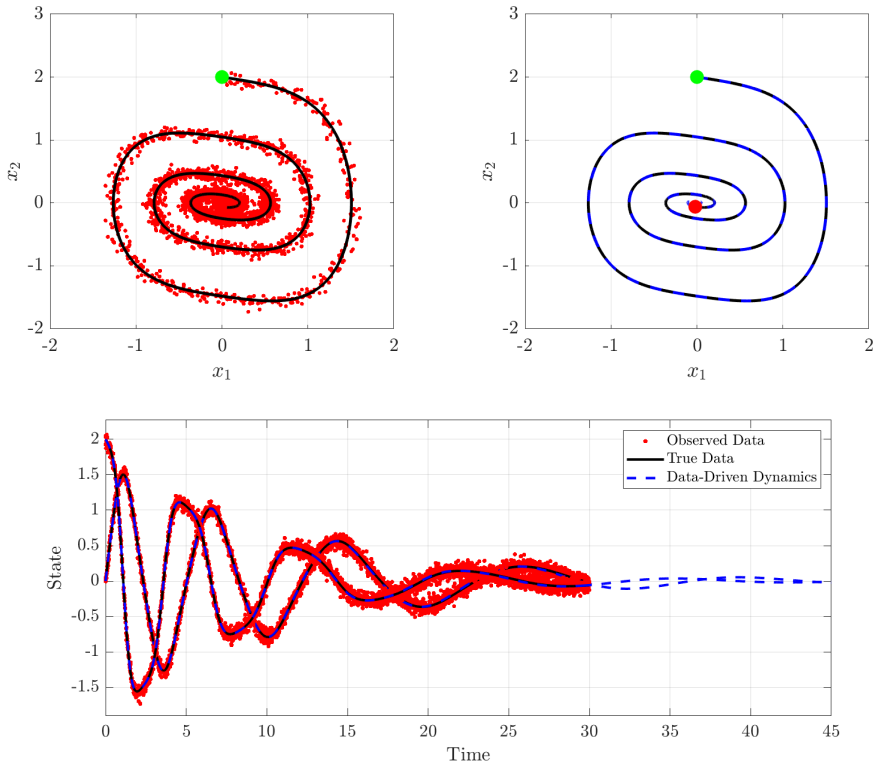


FIG. 3.5. *Large-noise regime: Duffing equation, $\beta = 1$. All correct terms were identified with an error in the weights of $E_2(\widehat{\mathbf{w}}) = 0.0075$ and a trajectory error of $\mathcal{E}_2(\widehat{\mathbf{w}}) = 0.014$.*

with trajectory error $\mathcal{E}_2(\widehat{\mathbf{w}}) < 0.02$. For the van der Pol oscillator and nonlinear pendulum (Figures 3.6 and 3.8), $\mathbf{x}_{dd}$ follows a limit cycle with an attractor that is indistiguishable from the true data (see phase plane plots); however, an error in the period of oscillation of roughly 0.6% leads to a larger trajectory error. The data-driven trajectory for the Lorenz equation diverges from the true trajectory around $t = 2.5$ (Figure 3.9), which is expected from chaotic dynamics, but still remains close to the Lorenz attractor.

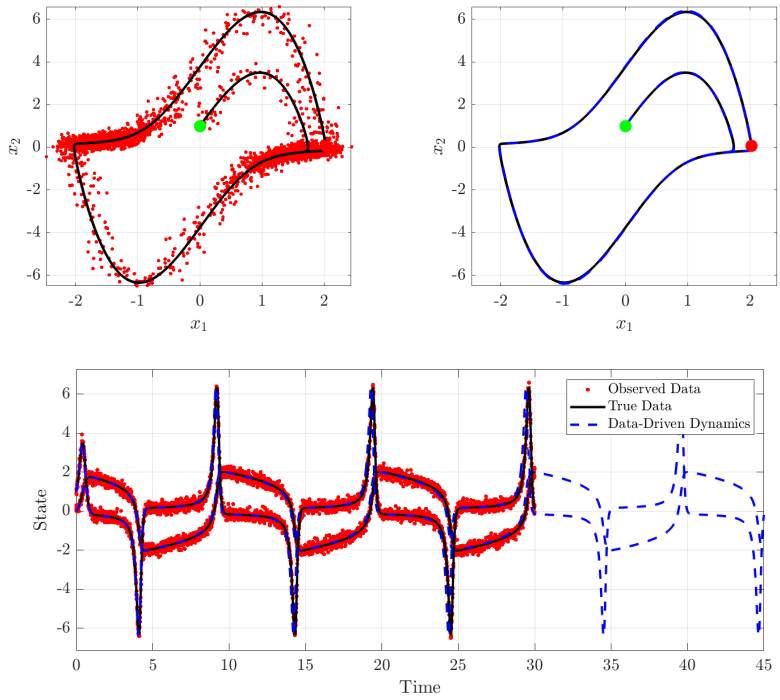


FIG. 3.6. *Large-noise regime: van der Pol oscillator,* $\beta = 4$. *All correct terms were identified with coefficient error* $E_2(\widehat{\mathbf{w}}) = 0.0073$ *and trajectory error* $\mathcal{E}_2(\widehat{\mathbf{w}}) = 0.32$. *The data-driven trajectory* $\mathbf{x}_{dd}$ *has a slightly shorter oscillation period of* 10.14 *time units compared to the true* 10.2, *resulting in an eventual offset from the true data* $\mathbf{x}$ *and hence a larger trajectory error. Measured over the time interval* $[0, 8]$ *the trajectory error is* 0.065.

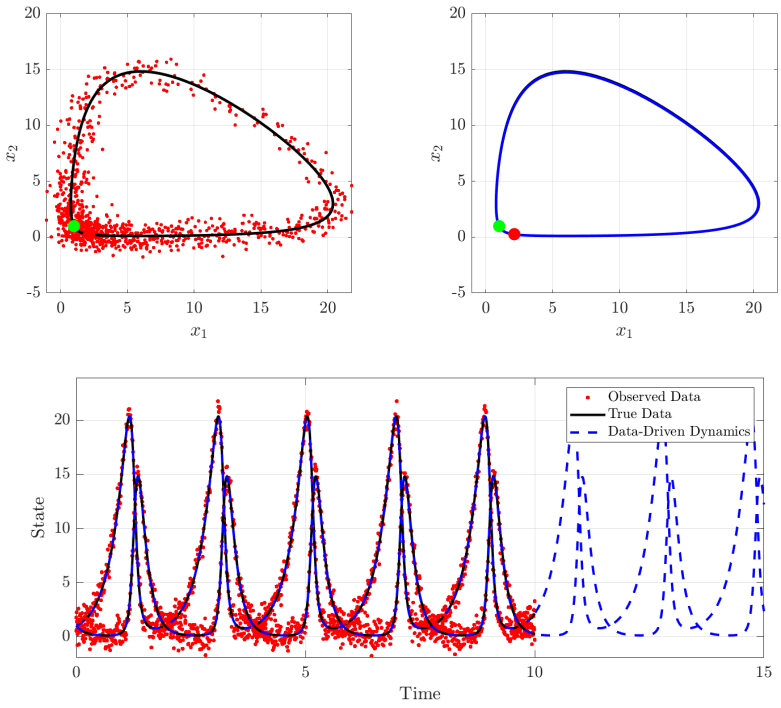


FIG. 3.7. *Large-noise regime: Lotka–Volterra system with* $\beta = 1$. *All correct nonzero terms were identified with an error in the weights of* $E_2(\widehat{\mathbf{w}}) = 0.0013$ *and trajectory error* $\mathcal{E}_2(\widehat{\mathbf{w}}) = 0.0082$.

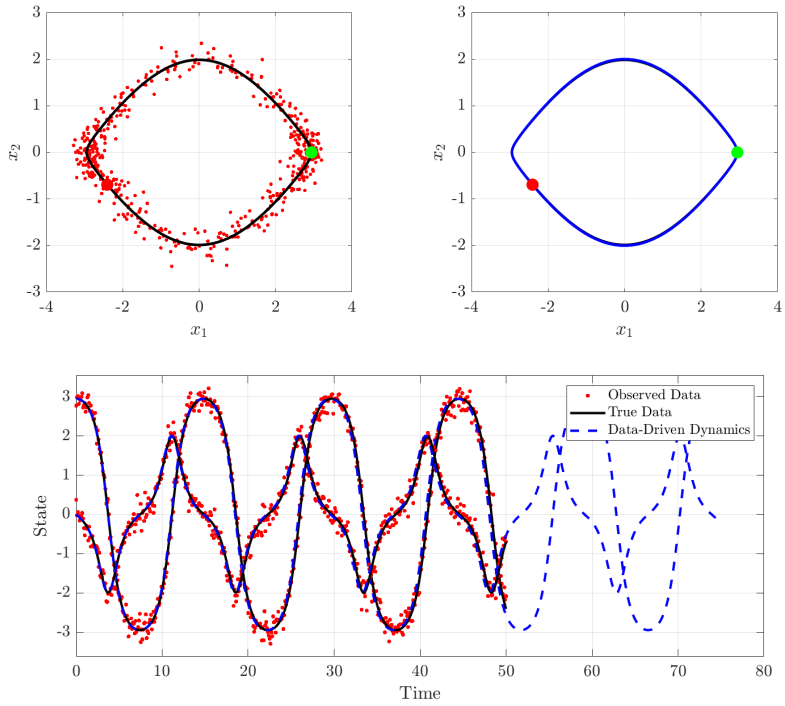


FIG. 3.8. *Large-noise regime: nonlinear pendulum with initial conditions* $\mathbf{x}(0) = (15\pi/16,\ 0)^T$. *All correct nonzero terms were identified with an error in the weights of* $E_2(\widehat{\mathbf{w}}) = 0.0089$ *and an error between* $\mathcal{E}_2(\widehat{\mathbf{w}}) = 0.076$.

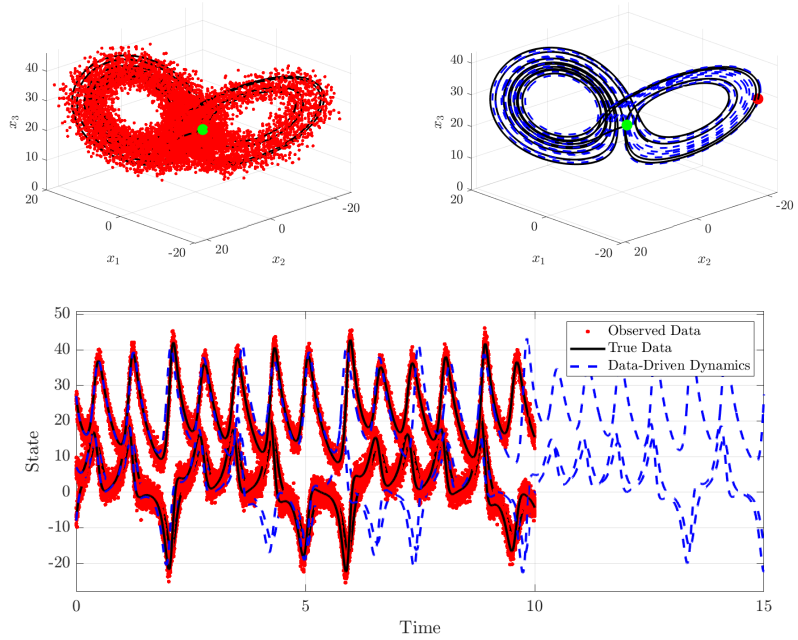


FIG. 3.9. *Large-noise regime: Lorenz system with* $\mathbf{x}_0 = (-8,\ 7,\ 27)^T$. *All correct terms were identified with an error in the weights of* $E_2(\widehat{\mathbf{w}}) = 0.0084$ *and trajectory error* $\mathcal{E}(\widehat{\mathbf{w}}) = 0.56$. *The large trajectory error is expected due to the chaotic nature of the solution. Using data up until* $t = 1.5$ *(first* 1500 *timepoints) the trajectory error is* 0.027.

4. Concluding remarks. We have developed and investigated a data-driven model selection algorithm based on the weak formulation of differential equations. The algorithm utilizes the reformulation of the model selection problem as a sparse regression problem for the weights $\mathbf{w}^\star$ of a candidate function basis $(f_j)_{j\in[J]}$ introduced in [21] and generalized in [4] as the SINDy algorithm. Our WSINDy algorithm can be seen as a generalization of the sparse recovery scheme using integral terms found in [17], where dynamics were recovered from noisy data using the integral equation. We have shown that by extending the integral equation to the weak form and using test functions with certain localization and smoothness properties, one may discovery the dynamics over a wide range of noise levels, with accuracy scaling favorably with noise: $E_2(\widehat{\mathbf{w}}) \approx 0.1\sigma_{NR}$.

A natural line of inquiry is to consider how WSINDy compares with conventional SINDy. There are several notable advantages of WSINDy; in particular, by considering the weak form of the equations, WSINDy completely avoids approximation of pointwise derivatives which significantly reduce the accuracy in conventional SINDy. When using SINDy, one must choose an appropriate numerical differentiation scheme depending on the noise level (e.g., finite differences are not robust to large noise but work well for small noise). For WSINDy, test functions from the space $\mathcal{S}$ (see section 2.4) together with the trapezoidal rule are effective in both low-noise and high-noise regimes. We demonstrate these observations in Appendix A by comparing WSINDy to SINDy under several numerical differentiation schemes. On the other hand, it may be the case that less data is required by standard SINDy. For the examples shown here, WSINDy works optimally for test functions supported on at least 15 timepoints, while many derivative approximations require fewer consecutive points.

WSINDy also utilizes the linearity of inner products with test functions to estimate the covariance structure of the residual, performing model selection in a generalized least squares framework. This is a much more appropriate setting given that residuals are neither independent nor uniformly distributed; however, we note that our implementations in this article employ approximate covariance matrices and could benefit from further refinements and investigation. In Appendix B we show that using generalized least squares with approximate covariance improves some results over ordinary least squares, but not significantly. We leave incorporation of more detailed knowledge of the covariance structure to future work. In addition, generalized least squares could potentially improve traditional model selection algorithms that rely on pointwise derivative estimates by similarly exploiting linear operators. Ultimately, a thorough analysis of the advantages of generalized least squares for model selection deserves further study.

Lastly, the most obvious extensions lie in generalizing the WSINDy method to spatiotemporal datasets. WSINDy as presented here in the context of ODEs is an exciting proof of concept with natural extensions to spatiotemporal and multiresolution settings building upon the extensive results in numerical and functional analysis for weak and variational formulations of physical problems.

Appendix A. Comparison between WSINDy and SINDy. Here we compare WSINDy and SINDy using the van der Pol oscillator, Lotka–Volterra system, and Lorenz equation. For WSINDy we place test functions along the time axis according to the uniform grid strategy. For SINDy, we examine three differentiation methods: total variation regularized derivatives (SINDy-TV), centered second-order finite difference (SINDy-FD-2), and centered fourth-order finite difference (SINDy-FD-4). For

SINDy-TV we use default settings and set the regularization parameter equal to the time step.

For each system and noise level we generate 200 independent instantiations of noise and record the average coefficient error $E_2(\widehat{\mathbf{w}})$ (3.2) as well as the average true positivity ratio (TPR) [10]:

$$\text{TPR}(\widehat{\mathbf{w}}) = \frac{\text{TP}(\widehat{\mathbf{w}})}{\text{TP}(\widehat{\mathbf{w}}) + \text{FP}(\widehat{\mathbf{w}}) + \text{FN}(\widehat{\mathbf{w}})}, \tag{A.1}$$

where $\text{TP}(\widehat{\mathbf{w}})$ is the number of correctly identified nonzero terms, $\text{FP}(\widehat{\mathbf{w}})$ is the number of falsely identified nonzero terms, and $\text{FN}(\widehat{\mathbf{w}})$ is the number of terms that are falsely identified as having a coefficient of zero. Since the feasible range of sparsity thresholds λ depends on the noise level, we adopt the selection methodology in [14] to choose an appropriate λ value for each instantiation of noise: λ is chosen from the set $10^{\{-5+\frac{i}{10}, i\in\{0,\dots,50\}\}}$ (i.e., the 51 values from 10^{-5} to 1 equally spaced $\log_{10}$) as the minimizer of the loss function

$$\mathcal{L}(\lambda) = \frac{\left\|\mathbf{A}\mathbf{w}^{\lambda} - \mathbf{A}\mathbf{w}^{0}\right\|_2}{\left\|\mathbf{A}\mathbf{w}^{0}\right\|_2} + \frac{\#\{j : \mathbf{w}_j^{\lambda} \neq 0\}}{J},$$

where $\mathbf{A} = \mathbf{V}\Theta(\mathbf{y})$ for WSINDy and $\mathbf{A} = \Theta(\mathbf{y})$ for SINDy; $\mathbf{w}^{\lambda}$ is the sequential-thresholding least squares solution for sparsity threshold λ, and J is the number of terms in the model library (for further details see [14]).

From Figures A.1, A.2, and A.3 we observe that for small noise (up to $\sigma_{NR} = 10^{-1}$), the coefficient error for WSINDy follows the linear trend $E_2(\widehat{\mathbf{w}}) \approx 0.1\sigma_{NR}$

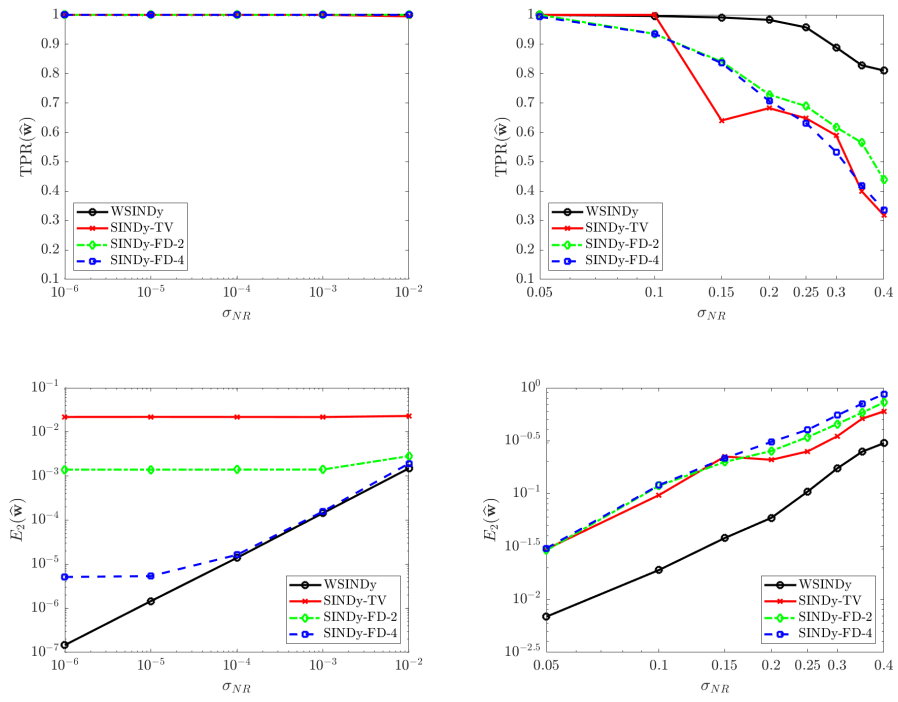


FIG. A.1. *Comparison between WSINDy and SINDy: van der Pol. Clockwise from top left: small-noise TPR($\widehat{\mathbf{w}}$) (defined in* (A.1)*), large-noise TPR($\widehat{\mathbf{w}}$), large-noise $E_2(\widehat{\mathbf{w}})$ (defined* (3.2)*), small-noise $E_2(\widehat{\mathbf{w}})$.*

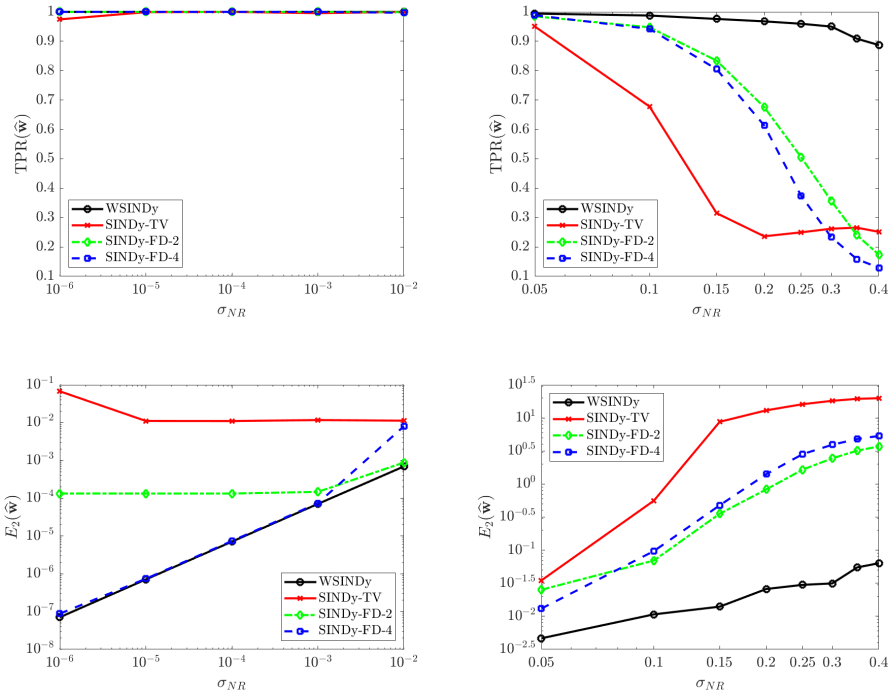


FIG. A.2. *Comparison between WSINDy and SINDy: Lotka–Volterra. Clockwise from top left: small-noise TPR($\widehat{\mathbf{w}}$) (defined in* (A.1)*), large-noise TPR($\widehat{\mathbf{w}}$), large-noise $E_2(\widehat{\mathbf{w}})$ (defined* (3.2)*), small-noise $E_2(\widehat{\mathbf{w}})$.*

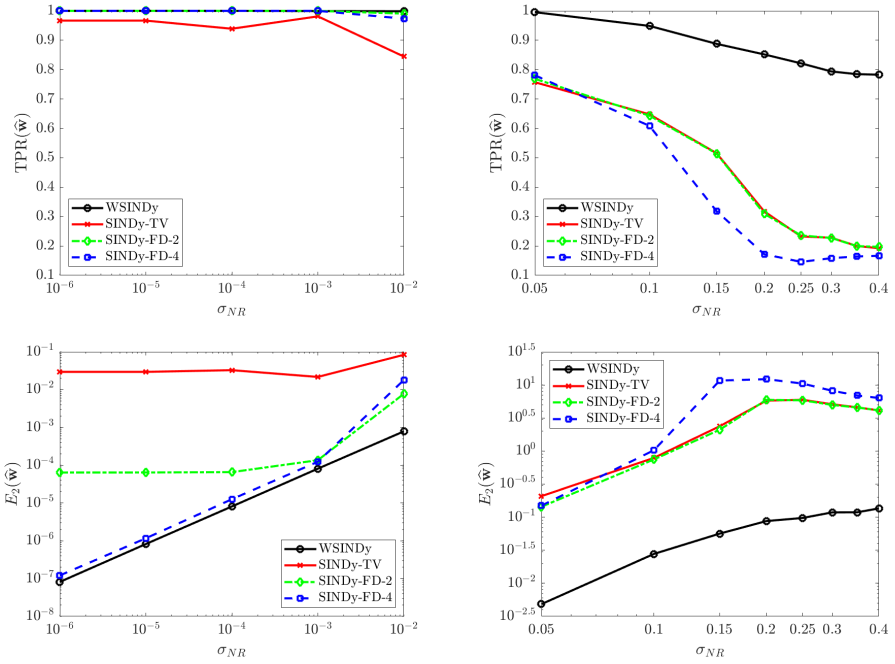


FIG. A.3. *Comparison between WSINDy and SINDy: Lorenz system. Clockwise from top left: small-noise TPR($\widehat{\mathbf{w}}$) (defined in* (A.1)*), large-noise TPR($\widehat{\mathbf{w}}$), large-noise $E_2(\widehat{\mathbf{w}})$ (defined* (3.2)*), small-noise $E_2(\widehat{\mathbf{w}})$.*

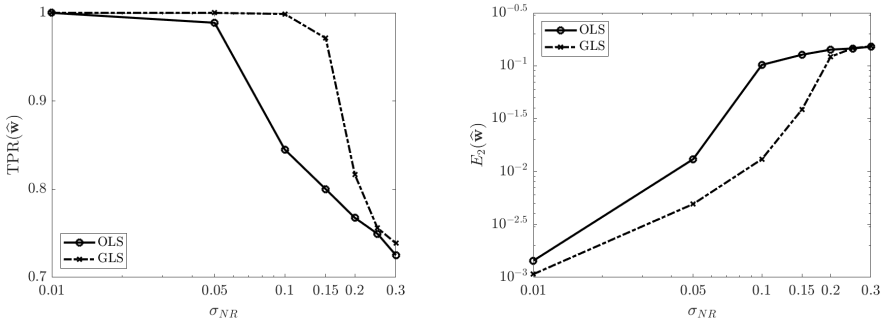


FIG. B.1. *Comparison between WSINDy with GLS and WSINDy with ordinary least squares using the Duffing equation. Results are averaged over* 200 *instantiations of noise.*

(observed in the text) and that SINDy-FD-4 behaves similarly but with slightly worse accuracy. For larger noise, SINDy diverges in accuracy and identification of the correct nonzero terms for each differentiation scheme, while WSINDy maintains a TPR of at least 0.8 up to 40% noise for each system. WSINDy thus provides an advantage across the entire noise spectrum examined, all while employing the same weak discretization scheme.

Appendix B. Generalized least squares vs. ordinary least squares. Generalized least squares (GLS) aims to account for correlations between the residuals [8]. Given a linear model $y = \mathbf{X}\beta + \epsilon$, where $\mathrm{Cov}(\epsilon) = \Sigma$ and $\mathbb{E}[\epsilon|\mathbf{X}] = 0$, the GLS estimator of the parameters β upon observing $\hat{y}$ is

$$\widehat{\beta} = \left(\mathbf{X}^T\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\Sigma^{-1}\hat{y}.$$

This provides the best linear unbiased estimator of β in the sense that if $\widetilde{\beta}$ is any other unbiased estimator, then $\widehat{\beta}$ has lower variance: $\mathbb{V}[\widehat{\beta}_i] \leq \mathbb{V}[\widetilde{\beta}_i]$, $i = 1, \ldots, n$.

Above we derived an approximate covariance matrix $\Sigma \approx \mathbf{V}'(\mathbf{V}')^T$ to use in the GLS implementation of WSINDy, although the true covariance depends on the underlying unknown dynamical system and hence is unattainable. In addition, since in our case $\mathbf{X} = \mathbf{G} = \mathbf{V}\Theta(\mathbf{y})$ depends on the noise ϵ, the assumption $\mathbb{E}[\epsilon|\mathbf{X}] = 0$ is violated. Nevertheless, we find that the noise regime $\sigma_{NR} \in [\,0.01, 0.3]$ does benefit from using GLS over ordinary least squares. Figure B.1 shows that for the Duffing equation, GLS extends the region $\{\sigma_{NR} \mid \mathrm{TPR}(\widehat{\mathbf{w}}) > 0.95\}$ from $\sigma_{NR} \leq 0.05$ to $\sigma_{NR} \leq 0.15$, as well as increases the accuracy in the recovered coefficients. This suggests that further improvements can be made with a more refined covariance matrix.

Acknowledgments. Code used in this manuscript is publicly available on GitHub at https://github.com/MathBioCU/WSINDy. The authors would like to thank Prof. Vanja Dukic (University of Colorado at Boulder, Department of Applied Mathematics) and Kadierdan Kaheman (University of Washington at Seattle, Department of Applied Mathematics) for helpful discussions.

REFERENCES

[1] H. AKAIKE, *A new look at the statistical model identification*, IEEE Trans. Automat. Control, 19 (1974), pp. 716–723, https://doi.org/10.1109/TAC.1974.1100705.

[2] H. AKAIKE, *On entropy maximization principle*, in Applications of Statistics, P. R. Krishnaiah, ed., North-Holland, Amsterdam, 1977, pp. 27–41

[3] D. M. Bortz and P. W. Nelson, *Model selection and mixed-effects modeling of HIV infection dynamics*, Bull. Math. Biol., 68 (2006), pp. 2005–2025, https://doi.org/10.1007/s11538-006-9084-x.
[4] 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. USA, 113 (2016), pp. 3932–3937.
[5] A. Cortiella, K.-C. Park, and A. Doostan, *Sparse identification of nonlinear dynamical systems via reweighted ℓ_1-regularized least squares*, Comput. Methods Appl. Mech. Engrg., 376 (2021), p. 113620.
[6] G. Dahlquist and A. Björck, *Numerical Methods in Scientific Computing: Volume* 1, vol. 103, SIAM, 2008.
[7] S. H. Kang, W. Liao, and Y. Liu, *IDENT: Identifying differential equations with numerical time evolution*, J. Sci. Comput., 87 (2021), 1.
[8] T. Kariya and H. Kurata, *Generalized Least Squares*, John Wiley & Sons, New York, 2004.
[9] R. T. Keller and Q. Du, *Discovery of dynamics using linear multistep methods*, SIAM J. Numer. Anal., 59 (2021), pp. 429–455.
[10] J. Lagergren, J. T. Nardini, G. M. Lavigne, E. M. Rutter, and K. B. Flores, *Learning partial differential equations for biological transport models from noisy spatio-temporal data*, Proc. A, 476 (2020), 20190800.
[11] J. H. Lagergren, J. T. Nardini, G. Michael Lavigne, E. M. Rutter, and K. B. Flores, *Learning partial differential equations for biological transport models from noisy spatio-temporal data*, Proc. A., 476 (2020), 20190800, https://doi.org/10.1098/rspa.2019.0800.
[12] G. Lillacci and M. Khammash, *Parameter estimation and model selection in computational biology*, PLoS Comput. Biol., 6 (2010), e1000696, https://doi.org/10.1371/journal.pcbi.1000696.
[13] F. Lu, M. Maggioni, and S. Tang, *Learning interaction kernels in heterogeneous systems of agents from multiple trajectories*, J. Mach. Learn. Res., 22 (2021), pp. 1–67.
[14] D. A. Messenger and D. M. Bortz, *Weak SINDy for Partial Differential Equations*, arXiv preprint, arXiv:2007.02848 [math.NA], 2020, https://arxiv.org/abs/2007.02848.
[15] M. Raissi, P. Perdikaris, and G. E. Karniadakis, *Machine learning of linear differential equations using Gaussian processes*, J. Comput. Phys., 348 (2017), pp. 683–693.
[16] S. H. Rudy, J. N. Kutz, and S. L. Brunton, *Deep learning of dynamics and signal-noise decomposition with time-stepping constraints*, J. Comput. Phys., 396 (2019), pp. 483–506.
[17] H. Schaeffer and S. G. McCalla, *Sparse model selection via integral terms*, Phys. Rev. E, 96 (2017), 023302.
[18] H. Schaeffer, G. Tran, R. Ward, and L. Zhang, *Extracting structured dynamical systems using sparse optimization with very few samples*, Multiscale Model. Simul., 18 (2020), pp. 1435–1461.
[19] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf, *Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems*, J. R. Soc. Interface., 6 (2009), pp. 187–202, https://doi.org/10.1098/rsif.2008.0172.
[20] G. Tran and R. Ward, *Exact recovery of chaotic systems from highly corrupted data*, Multiscale Model. Simul., 15 (2017), pp. 1108–1129.
[21] 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 (2011), p. 154101.
[22] D. J. Warne, R. E. Baker, and M. J. Simpson, *Using experimental data and information criteria to guide model selection for reaction–diffusion problems in mathematical biology*, Bull. Math. Biol., 81 (2019), pp. 1760–1804, https://doi.org/10.1007/s11538-019-00589-x.
[23] H. Wu and L. Wu, *Identification of significant host factors for HIV dynamics modelled by non-linear mixed-effects models*, Stat. Med., 21 (2002), pp. 753–771, https://doi.org/10.1002/sim.1015.
[24] K. Wu, T. Qin, and D. Xiu, *Structure-preserving method for reconstructing unknown Hamiltonian systems from trajectory data*, SIAM J. Sci. Comput., 42 (2020), pp. A3704–A3729.
[25] K. Wu and D. Xiu, *Numerical aspects for approximating governing equations using data*, J. Comput. Phys., 384 (2019), pp. 200–221.
[26] S. Zhang and G. Lin, *Robust data-driven discovery of governing physical laws with error bars*, Proc. A, 474 (2018), 20180305.
[27] S. Zhang and G. Lin, *Robust Subsampling-Based Sparse Bayesian Inference to Tackle Four Challenges (Large Noise, Outliers, Data Integration, and Extrapolation) in the Discovery of Physical Laws from data*, arXiv preprint, arXiv:1907.07788 [stat.ML], 2019, https://arxiv.org/abs/1907.07788.