



# Direct Estimation of Parameters in ODE Models Using WENDy: Weak-Form Estimation of Nonlinear Dynamics

David M. Bortz<sup>1</sup>  · Daniel A. Messenger<sup>1</sup> · Vanja Dukic<sup>1</sup>

Received: 26 February 2023 / Accepted: 30 August 2023

© The Author(s) 2023

## Abstract

We introduce the Weak-form Estimation of Nonlinear Dynamics (WENDy) method for estimating model parameters for non-linear systems of ODEs. Without relying on any numerical differential equation solvers, WENDy computes accurate estimates and is robust to large (biologically relevant) levels of measurement noise. For low dimensional systems with modest amounts of data, WENDy is competitive with conventional forward solver-based nonlinear least squares methods in terms of speed and accuracy. For both higher dimensional systems and stiff systems, WENDy is typically both faster (often by orders of magnitude) and more accurate than forward solver-based approaches. The core mathematical idea involves an efficient conversion of the strong form representation of a model to its weak form, and then solving a regression problem to perform parameter inference. The core statistical idea rests on the Errors-In-Variables framework, which necessitates the use of the iteratively reweighted least squares algorithm. Further improvements are obtained by using orthonormal test functions, created from a set of  $C^\infty$  bump functions of varying support sizes. We demonstrate the high robustness and computational efficiency by applying WENDy to estimate parameters in some common models from population biology, neuroscience, and biochemistry, including logistic growth, Lotka-Volterra, FitzHugh-Nagumo, Hindmarsh-Rose, and a Protein Transduction Benchmark model. Software and code for reproducing the examples is available at <https://github.com/MathBioCU/WENDy>.

**Keywords** Data-driven modeling · Parameter estimation · Parameter inference · Weak Form · Test functions

---

This research was supported in part by the following Grants: NSF Mathematical Biology MODULUS Grant 2054085 to DMB; NSF/NIH Joint DMS/NIGMS Mathematical Biology Initiative Grant R01GM126559 to DMB; DOE ASCR MMICC Grant DE-SC0023346 to DMB; NIFA Biological Sciences Grant 2019-67014-29919 to VD; and NSF Division Of Environmental Biology Grant 2109774 to VD. Software and code for reproducing the examples is available at <https://github.com/MathBioCU/WENDy>.

---

✉ David M. Bortz  
david.bortz@colorado.edu

<sup>1</sup> Department of Applied Mathematics, University of Colorado, Boulder CO 80309-0526, USA

**Mathematics Subject Classification** 35D30 · 62FXX · 62JXX · 65L09 · 65M32 · 92-08

## 1 Introduction

Accurate estimation of parameters for a given model is central to modern scientific discovery. It is particularly important in the modeling of biological systems which can involve both first principles-based and phenomenological models and for which measurement errors can be substantial, often in excess of 20%. The dominant methodologies for parameter inference are either not capable of handling realistic errors, or are computationally costly relying on forward solvers or Markov chain Monte Carlo methods. In this work, we propose an accurate, robust and efficient weak form-based approach to estimate parameters for parameter inference. We demonstrate that our “Weak form Estimation of Nonlinear Dynamics” (WENDy) method offers many advantages including high accuracy, robustness to substantial noise, and computational efficiency often up to several orders of magnitude over the existing methods.

In the remainder of this section, we provide an overview of modern parameter estimation methods in ODE systems, as well as a discussion of the literature that led to the WENDy idea. Section 2 contains the core weak-form estimation ideas as well as the WENDy algorithm itself. In Sect. 2.1, we introduce the idea of weak-form parameter estimation, including a simple algorithm to illustrate the idea. In Sect. 2.2, we describe the WENDy method in detail. We describe the Errors-In-Variables (EiV) framework, and derive a Taylor expansion of the residual which allows us to formulate the (in Sect. 2.2) Iteratively Reweighted Least Squares (IRLS) approach to inference. The EiV and IRLS modifications are important as they offers significant improvements to the Ordinary Least Squares approach. In Sect. 2.3, we present a strategy for computing an orthogonal set of test functions that facilitate a successful weak-form implementation. In Sect. 3 we illustrate the performance of WENDy using five common mathematical models from the biological sciences and in Sect. 4 we offer some concluding remarks.

### 1.1 Background

A ubiquitous version of the parameter estimation problem in the biological sciences is

$$\hat{\mathbf{w}} := \arg \min_{\mathbf{w} \in \mathbb{R}^J} \|u(\mathbf{t}; \mathbf{w}) - \mathbf{U}\|_2^2, \quad (1)$$

where the function  $u : \mathbb{R} \rightarrow \mathbb{R}^d$  is a solution to a differential equation model<sup>1</sup>

$$\begin{aligned} \dot{u} &= \sum_{j=1}^J w_j f_j(u), \\ u(t_0) &= u_0 \in \mathbb{R}^d, \end{aligned} \quad (2)$$

<sup>1</sup> While we restrict ourselves to deterministic differential equations, there is nothing in the WENDy approach that inhibits extension to discrete or stochastic models.

The ODE system in (2) is parameterized by  $\mathbf{w} \in \mathbb{R}^J$ , the vector of  $J$  true parameters which are to be estimated by  $\hat{\mathbf{w}}$ . The solution to the equation is then compared (in a least squares sense) with data  $\mathbf{U} \in \mathbb{R}^{(M+1) \times d}$  that is sampled at  $M + 1$  timepoints  $t := \{t_i\}_{i=0}^M$ . We note that in this work, we will restrict the differential equations to those with right sides that are linear combinations of the  $f_j$  functions with coefficients  $w_j$ , as in Eq. (2).

Conventionally, the standard approach for parameter estimation methodologies has been forward solver-based nonlinear least squares (FSNLS). In that framework, 1) a candidate parameter vector is proposed, 2) the resulting equation is numerically solved on a computer, 3) the output is compared (via least squares) to data, and 4) then this process is repeated until a convergence criteria is met. This is a mature field and we direct the interested reader to references by Ljung (1999, 2017) and, for those interested in a more theoretical perspective, to the monograph by Banks and Kunisch (1989).

The FSNLS methodology is very well understood and its use is ubiquitous in the biological, medical, and bioengineering sciences. However, as models get larger and more realism is demanded of them, there remain several important challenges that do not have fully satisfying answers. For example, the accuracy of the solver can have a huge impact on parameter estimates; see (Nardini and Bortz 2019) for an illustration with PDE models and Bortz (2006) for an example with ODE and DDE models. There is no widespread agreement on a method to detect this type of error and the conventional strategy would be to simply increase the solution accuracy (usually at significant computational cost) until the estimate stabilizes.

Given the above, it is reasonable to consider alternatives to fitting via comparing an approximate model solution with the measured data. A natural idea would be to avoid performing forward solves altogether via substituting the data directly into the model Eq. (2). The derivative could be approximated via differentiating a projection of the data onto, e.g., orthogonal polynomials, and the parameters could then be estimated by minimizing the norm of the residual of the Eq. (2)—i.e., via a gradient matching criteria. Indeed, Richard Bellman proposed exactly this strategy in 1969 (Bellman 1969). There have been similar ideas in the literature of chemical and aerospace engineering, which can be traced back even further (Perdreauville and Goodson 1966; Greenberg 1951). However, these methods are known to perform poorly in the presence of even modest noise.

To account for the noise in the measurements while estimating the parameters (and in some cases the state trajectories), researchers have proposed a variety of different non-solver-based methods. The most popular modern approaches involve denoising the measured state via Gaussian Processes (Yang et al. 2021; Martina-Perez et al. 2021; Wang and Zhou 2021; Wenk et al. 2020; Calderhead et al. 2008) and collocations projecting onto a polynomial or spline basis (Varah 1982; Ramsay et al. 2007; Liang and Wu 2008; Poyton et al. 2006; Brunel 2008; Zhang et al. 2022). For example, Yang et al. (2021), restricted a Gaussian Process to the manifold of solutions to an ODE to infer both the parameters and the state using a Hamiltonian Markov chain Monte Carlo method. Ramsay et al. (2007) proposed a collocation-type method in which the solution is projected onto a spline basis. In a two-step procedure, both the basis weights and the unknown parameters are iteratively estimated. The minimization

identifies the states and the parameters by penalizing poor faithfulness to the model equation (i.e., gradient matching) and deviations too far from the measured data. Liang and Wu (2008) proposed a similar strategy based on local polynomial smoothing to first estimate the state and its derivative, compute derivatives of the smoothed solution, and then estimate the parameters. Ding and Wu later improved upon this work in Ding and Wu (2014) by using local polynomial regression instead of the pseudo-least squares estimator used in Liang and Wu (2008).

There are also a few approaches which focus on transforming the equations with operators that allow efficiently solving for the parameters. In particular Xu and Khanmohamadi created smoothing and derivative smoothing operators based on Fourier theory (Xu et al. 2008) and Chebyshev operators (Khanmohamadi and Xu 2009). However, they have not proven to be as influential as the integral and weak form methods described in the next subsection.

## 1.2 Integral and Weak Form Methods

Recent efforts by our group and others suggest that there is a considerable advantage in parameter estimation performance to be gained from using an integral-based transform of the model equations. The two main approaches are to (1) use integral forms of the model equation or (2) convolve the equation with a compactly supported test function to obtain the so-called "weak form" of the equation. The weak form idea can be traced back to Laurent Schwartz's Theory of Distributions (Schwartz 1950),<sup>2</sup> which recasts the classical notion of a function acting on a point to one acting on a measurement structure or "test function". In the context of differential equation models, Lax and Milgram pioneered the use of the weak form for relaxing smoothness requirements on unique solutions to parabolic PDE systems in Hilbert spaces (Lax and Milgram 1955). Since then, the weak form has been heavily used in studying solutions to PDEs as well as numerically solving for the solutions (e.g., the Finite Element Method), but not with the goal of directly estimating parameters.

The idea of weak-form based estimation has been repeatedly discovered over the years (see (Preisig and Rippin 1993) for a good historical overview). Briefly, in 1954, Shinbrot created a proto-weak-form parameter inference method, called the Equations Of Motion (EOM) method (Shinbrot 1954). In it, he proposes to multiply the model equations by so-called method functions, i.e., what we would now call test functions. These test functions were based on  $\sin^n(\nu t)$  for different values of  $\nu$  and  $n$ . In 1965, Loeb and Cahen (1965a,b) independently discovered the same method, calling it the Modulating Function (MF) method. They proposed and advocated for the use of polynomial test functions. The issue with these approaches (and indeed all subsequent developments based on these methods) is that the maximum power  $n$  is chosen to exactly match the number of derivatives needed to perform integration by parts (IBP). As we have shown, this choice means that these methods are not nearly as effective as they could be. As we initially reported in Messenger and Bortz (2021b), a critical step in obtaining robust and accurate parameter estimation is to use *highly* smooth test

<sup>2</sup> See (Duistermaat and Kolk 2010) for a modern introduction.

functions, e.g., to have  $n$  be substantially higher than the minimum needed by the IBP. This insight led to our use of the  $C^\infty$  bump functions in WENDy (see Sect. 2.3).

In the statistics literature, there are several examples of using integral or weak-form equations. Dattner et al. (2017) illustrate an integral-based approach and Dattner's 2021 review (Dattner 2021) provides a good overview of other efforts to use the integral form for parameter estimation. Concerning the weak form, several researchers have used it as a core part of their estimation methods (see works by Brunel et al. 2014 and Sangalli 2021). Unlike WENDy, however, either these approaches smooth the data before substitution into the model equation (which can lead to poor performance) or still require forward solves. As with the EOM and MF method above, the test functions in these methods were also chosen with insufficient smoothness to yield the highly robust parameter estimates we obtain with WENDy.

As the field of SINDy-based equation learning (Brunton et al. 2016) is built upon direct parameter estimation methods, there are also several relevant contributions from this literature. Schaeffer and McCalla (2017) showed that parameter estimation and learning an integral form of equations can be done in the presence of significant noise. Broadly speaking, however, the consensus has emerged that the weak form is more effective than a straightforward integral representation. In particular, several groups (including ours) independently proposed weak form-based approaches (Pantazis and Tsamardinou 2019; Gurevich et al. 2019; Messenger and Bortz 2021b; Wang et al. 2019; Messenger and Bortz 2021a). The weak form is now even implemented in the PySINDy code (Kaptanoglu et al. 2022) which is actively developed by the authors of the original SINDy papers (Brunton et al. 2016; Rudy et al. 2017). However, we do note that the Weak SINDy in PySINDy is based on an early weak form implementation (proposed in Gurevich et al. 2019; Reinbold et al. 2020). A more recent implementation with autotuned hyperparameters can be found at [https://github.com/MathBioCU/WSINDy\\_ODE](https://github.com/MathBioCU/WSINDy_ODE) for ODEs (Messenger and Bortz 2021b) and [https://github.com/MathBioCU/WSINDy\\_PDE](https://github.com/MathBioCU/WSINDy_PDE) for PDEs (Messenger and Bortz 2021a).

While our group wasn't the first to propose a weak form methodology, we have pioneered its use for equation learning in a wide range of model structures and applications including: ODEs (Messenger and Bortz 2021b), PDEs (Messenger and Bortz 2021a), interacting particle systems of the first (Messenger and Bortz 2022b) and second (Messenger et al. 2022b) order, and online streaming (Messenger et al. 2022a). We have also studied and advanced the computational method itself. Among other contributions, we were the first to automate (with mathematical justification) test function hyperparameter specification, feature matrix rescaling (to ensure stable computations), and to filter high frequency noise (Messenger and Bortz 2021a). Lastly we have also studied the theoretical convergence properties for WSINDy in the continuum data limit (Messenger and Bortz 2022a). Among the results are a description of a broad class of models for which the asymptotic limit of continuum data can overcome *any* noise level to yield both an accurately learned equation and a correct parameter estimate (see Messenger and Bortz 2022a for more information).

## 2 Weak form Estimation of Nonlinear Dynamics (WENDy)

In this work, we assume that the exact form of a differential equation-based mathematical model is known, but that the precise values of constituent parameters are to be estimated using existing data. As the model equation is not being learned, this is different than the WSINDy methodology and, importantly, does not use sparse regression. We thus denote the method presented in this paper as the Weak-form Estimation of Nonlinear Dynamics (WENDy) method.

In Sect. 2.1, we start with an introduction to the idea of weak-form parameter estimation in a simple OLS setting. In Sect. 2.2 we describe the WENDy algorithm in detail, along with several strategies for improving the accuracy: in Sect. 2.3 we describe a strategy for optimal test function selection, and in Sect. 2.4 the strategy for improved iteration termination criteria.

### 2.1 Weak-Form Estimation with Ordinary Least Squares

We begin by considering a  $d$ -dimensional matrix form of (2), i.e., an ordinary differential equation system model

$$\dot{u} = \Theta(u)W \quad (3)$$

with row vector of the  $d$  solution states  $u(t; W) := [u_1(t; W) | u_2(t; W) | \cdots | u_d(t; W)]$ , row vector of  $J$  features (i.e., right side terms<sup>3</sup> where  $f_j : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $C_c^2$ ) such that  $\Theta(u) := [f_1(u) | f_2(u) | \cdots | f_J(u)]$ , and the matrix of unknown parameters  $W \in \mathbb{R}^{J \times d}$ . We consider a  $C^\infty$  test function  $\phi$  compactly supported in the time interval  $[0, T]$  (e.g.  $\phi \in C_c^\infty([0, T])$ ), multiply both sides of (3) by  $\phi$ , and integrate over 0 to  $T$ . Via integration by parts we obtain

$$\phi(T)u(T) - \phi(0)u(0) - \int_0^T \dot{\phi}u dt = \int_0^T \phi \Theta(u)W dt.$$

As the compact support of  $\phi$  implies that  $\phi(0) = \phi(T) = 0$ , this yields a transform of (3) into

$$- \int_0^T \dot{\phi}u dt = \int_0^T \phi \Theta(u)W dt. \quad (4)$$

This weak form of the equation allows us to define a novel methodology for estimating the entries in  $W$ .

Observations of states of this system are (in this paper) assumed to occur at a discrete set of  $M + 1$  timepoints  $\{t_m\}_{m=0}^M$  with uniform stepsize  $\Delta t$ . The test functions are thus centered at a subsequence of  $K$  timepoints  $\{t_{m_k}\}_{k=1}^K$ . We choose the test function support to be centered at a timepoint  $t_{m_k}$  with radius  $m_t \Delta t$  where  $m_t$  is an

<sup>3</sup> Note that here we assume that  $f_j$  is in  $C_c^2$ , simply to ensure that the Taylor expansion in Sect. 2.2 is well-defined.

integer (to be chosen later). Bold variables denote evaluation at or dependence on the chosen timepoints, e.g.,

$$\mathbf{t} := \begin{bmatrix} t_0 \\ \vdots \\ t_M \end{bmatrix}, \mathbf{u} := \begin{bmatrix} u_1(t_0) & \cdots & u_d(t_0) \\ \vdots & \ddots & \vdots \\ u_1(t_M) & \cdots & u_d(t_M) \end{bmatrix}, \Theta(\mathbf{u}) := \begin{bmatrix} f_1(u(t_0)) & \cdots & f_J(u(t_0)) \\ \vdots & \ddots & \vdots \\ f_1(u(t_M)) & \cdots & f_J(u(t_M)) \end{bmatrix}.$$

Approximating the integrals in (4) using a Newton-Cotes quadrature yields

$$-\dot{\phi}_k \mathbf{u} \approx \phi_k \Theta(\mathbf{u}) W, \quad (5)$$

where

$$\phi_k := [\phi_k(t_0) | \cdots | \phi_k(t_M)] \mathcal{Q}, \quad \dot{\phi}_k := [\dot{\phi}_k(t_0) | \cdots | \dot{\phi}_k(t_M)] \mathcal{Q}$$

and  $\phi_k$  is a test function centered at timepoint  $t_{m_k}$ . To account for proper scaling, in computations we normalize each test function  $\phi_k$  to have unit  $\ell_2$ -norm, or  $\sum_{m=0}^M \phi_k^2(t_m) = 1$ .

The  $\mathcal{Q}$  matrix contains the quadrature weights on the diagonal. In this work we use the composite Trapezoidal rule<sup>4</sup> for which the matrix is

$$\mathcal{Q} := \text{diag}(\Delta t/2, \Delta t, \dots, \Delta t, \Delta t/2) \in \mathbb{R}^{(M+1) \times (M+1)}.$$

We defer full consideration of the integration error until Sect. 2.3.1 but note that in the case of a non-uniform timegrid,  $\mathcal{Q}$  would simply be adapted with the correct stepsize and quadrature weights.

The core idea of the weak-form-based direct parameter estimation is to identify  $W$  as a least squares solution to

$$\min_W \|\text{vec}(\mathbf{G}W - \mathbf{B})\|_2^2 \quad (6)$$

where “vec” vectorizes a matrix,

$$\begin{aligned} \mathbf{G} &:= \phi \Theta(\mathbf{U}) \in \mathbb{R}^{K \times J}, \\ \mathbf{B} &:= -\dot{\phi} \mathbf{U} \in \mathbb{R}^{K \times d}, \end{aligned}$$

where  $\mathbf{U} \in \mathbb{R}^{(M+1) \times d}$  represents the data, and the integration matrices are

$$\phi = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_K \end{bmatrix} \in \mathbb{R}^{K \times (M+1)} \quad \text{and} \quad \dot{\phi} = \begin{bmatrix} \dot{\phi}_1 \\ \vdots \\ \dot{\phi}_K \end{bmatrix} \in \mathbb{R}^{K \times (M+1)}.$$

<sup>4</sup> The composite Trapezoidal rule works best for the uniform spacing and thus the left and right sides of (5) are sums weighted by  $\dot{\phi}_k(\mathbf{t})$  and  $\phi_k(\mathbf{t})$ , respectively.

In much of the previous work in regression-based data-driven modeling (including our own), the solution to the problem in (6) was computed by solving the normal equations to minimize the (Euclidean norm) residual (see Messenger and Bortz 2021a; Fasel et al. 2021; Nicolaou et al. 2023; Bertsimas and Gurnee 2023; Brunton et al. 2016 for examples). When the errors are present only in the output of the linear function (e.g., having errors only in  $\mathbf{b}$ ), and under the assumption that those errors are independent and identically distributed (i.i.d.) Gaussian random variables, this is known as the Ordinary Least Squares (OLS) method for solving linear least squares problems. In that case, as the number of data points increases, asymptotically the parameter estimate converges in probability to the true parameter (i.e., the OLS estimate is a *consistent* estimator).

The OLS solution to the regression problem in (6) is presented in Algorithm 1. We note that we have written the algorithm this way to promote clarity concerning the weak-form estimation idea. For actual implementation, we create a different  $\Theta_i$  for each variable  $i = 1 \dots, d$  and use regression for state  $i$  to solve for a vector  $\hat{\mathbf{w}}_i$  of parameters (instead of a matrix of parameters  $\mathbf{W}$ , which can contain values known to be zero). To increase computational efficiency, we make sure to remove any redundancies and use sparse computations whenever possible.

**Algorithm 1:** Weak-form Parameter Estimation with Ordinary Least Squares

**input** : Data  $\{\mathbf{U}\}$ , Feature Map  $\{\Theta\}$ , Test Function Matrices  $\{\phi, \dot{\phi}\}$   
**output**: Parameter Estimate  $\{\hat{\mathbf{W}}\}$

```
// Solve Ordinary Least Squares Problem
1  $\mathbf{G} \leftarrow \phi \Theta(\mathbf{U})$ 
2  $\mathbf{B} \leftarrow -\dot{\phi} \mathbf{U}$ 
3  $\hat{\mathbf{W}} \leftarrow (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{B}$ 
```

For this OLS problem, in order for the linear regression in Algorithm 1 to have a unique solution,  $\mathbf{G}$  must be full rank (i.e.,  $\text{rank}(\mathbf{G}) = J$ ) and we now present the conditions needed to satisfy this criteria. As we have done everywhere in this work, we assume that there are  $M + 1$  sample points in time with stepsize  $\Delta t$  and for each  $k$ ,  $\phi_k$  is centered at  $t_{m_k}$  and compactly supported on  $[t_k - m_t \Delta t, t_k + m_t \Delta t]$ . Thus, for  $J$  features,  $K$  test functions and  $M$  samples in time, there is a unique solution to the OLS problem when all of the following conditions hold.

**Condition 1**  $\phi_k \in C_c^p(\mathbb{R})$  for any  $p \geq 1$ .

**Condition 2**  $\text{rank}(\phi) = K$

**Condition 3**  $\text{rank}(\Theta) = J$

**Condition 4**  $J \leq K \leq M + 1 - 2m_t$ .

The first condition requires that all test functions  $\{\phi_k\}_{i=1}^K$  and their first derivatives  $\{\dot{\phi}_k\}_{i=1}^K$  have compact support so that the conversion to the weak form is valid. The



second condition means that the test functions must be chosen so that they are distinct. The third condition means that for each feature, in at least some sub-region of the sampled trajectory, the evaluation of that feature changes over time.<sup>5</sup> The last condition simply enforces the relationship between the cardinality of features, test functions (and their radii), and sampled points. All 4 of these conditions must be true to ensure that  $\mathbf{G}$  is full rank.

The OLS solution has respectable performance in some cases, but in general there is a clear need for improvement upon OLS.<sup>6</sup> In particular, we note that (6) is *not* a standard least squares problem. The (likely noisy) observations of the state  $u$  appear on both sides of (5). In Statistics, this is known as an Errors in Variables (EiV) problem.<sup>7</sup> While a full and rigorous analysis of the statistical properties of weak-form estimation is beyond the scope of this article,<sup>8</sup> here we will present several formal derivations aimed at improving the accuracy of weak-form parameter estimators. These improvements are critical as the OLS approach is not reliably accurate. Accordingly, we define WENDy (in the next section) as a weak-form parameter estimation method which uses techniques that address the EiV challenges.

## 2.2 WENDy: Weak-Form Estimation Using Iterative Reweighting

In this subsection, we address the fact that the posed regression problem does not fit within the framework of ordinary least squares, and is actually an Errors-In-Variables problem. We now derive a linearization that yields insight into the covariance structure of the problem. First, we denote the vector of true (but unknown) parameter values used in all state variable equations as  $\mathbf{w}^*$  and let  $u^* := u(t; \mathbf{w}^*)$  and  $\Theta^* := \Theta(u^*)$ . We also assume that measurements of the system are noisy, so that at each timepoint  $t$  all states are observed with additive noise

$$U(t) = u^*(t) + \varepsilon(t) \quad (7)$$

where each element of  $\varepsilon(t)$  is i.i.d.  $\mathcal{N}(0, \sigma^2)$ .<sup>9</sup> Lastly, we note that there are  $d$  variables,  $J$  feature terms, and  $M + 1$  timepoints. In what follows, we present the expansion using Kronecker products (denoted as  $\otimes$ ).

We begin by considering the sampled data  $\mathbf{U} := \mathbf{u}^* + \boldsymbol{\varepsilon} \in \mathbb{R}^{(M+1) \times d}$  and vector of parameters to be identified  $\mathbf{w} \in \mathbb{R}^{Jd}$ . We use bolded variables to represent evaluation

<sup>5</sup> In practice, this means that one must have enough data such that the columns of  $\Theta$  are linearly independent, which can easily be verified on any given data set. Identifying when this happens involves questions of data sufficiency and is beyond the scope of this work.

<sup>6</sup> See upper right plots in Figs. 5, 6, 7, 8, 9 and 9 for illustration of the improvements over the OLS method.

<sup>7</sup> Errors in Variables models are regression models accounting for errors in both the dependent and independent variables. If the ratio between the variances of dependent and independent variables is known, the 1D regression is called Demming regression while the multidimensional case is known as Total Least Squares (TLS). We direct the interested reader to Van et al. (2002) for more information.

<sup>8</sup> See our work in Messenger and Bortz (2022a) for an investigation of the asymptotic consistency in the limit of continuum data.

<sup>9</sup> Naturally, for real data, there could be different variances for different states as well as more sophisticated measurement error models. We defer such questions to future work.

at the timegrid  $\mathbf{t}$ , and use superscript  $\star$  notation to denote quantities based on true (noise-free) parameter or states. We now consider the residual

$$\mathbf{r}(\mathbf{U}, \mathbf{w}) := \mathbf{G}\mathbf{w} - \mathbf{b}, \quad (8)$$

where we redefine

$$\begin{aligned} \mathbf{G} &:= [\mathbb{I}_d \otimes (\phi \Theta(\mathbf{U}))], \\ \mathbf{b} &:= -\text{vec}(\dot{\phi}\mathbf{U}). \end{aligned}$$

We then note that we can decompose the residual into several components

$$\mathbf{r}(\mathbf{U}, \mathbf{w}) = \mathbf{G}\mathbf{w} - \mathbf{G}^\star\mathbf{w} + \mathbf{G}^\star\mathbf{w} - \mathbf{G}^\star\mathbf{w}^\star + \mathbf{G}^\star\mathbf{w}^\star - (\mathbf{b}^\star + \mathbf{b}^\varepsilon) \quad (9)$$

$$= \underbrace{(\mathbf{G} - \mathbf{G}^\star)\mathbf{w}}_{\mathbf{e}_\Theta} + \underbrace{\mathbf{G}^\star(\mathbf{w} - \mathbf{w}^\star)}_{\mathbf{r}_0} + \underbrace{(\mathbf{G}^\star\mathbf{w}^\star - \mathbf{b}^\star)}_{\mathbf{e}_{\text{int}}} - \mathbf{b}^\varepsilon, \quad (10)$$

where

$$\begin{aligned} \mathbf{G}^\star &:= [\mathbb{I}_d \otimes (\phi \Theta(\mathbf{u}^\star))], \\ \mathbf{b} &:= \underbrace{-\text{vec}(\dot{\phi}\mathbf{u}^\star)}_{\mathbf{b}^\star} + \underbrace{-\text{vec}(\dot{\phi}\boldsymbol{\varepsilon})}_{\mathbf{b}^\varepsilon}. \end{aligned}$$

Here,  $\mathbf{r}_0$  is the residual without measurement noise or integration errors, and  $\mathbf{e}_{\text{int}}$  is the numerical integration error induced by the quadrature (and will be analyzed in Sect. 2.3).

Let us further consider the leftover terms  $\mathbf{e}_\Theta - \mathbf{b}^\varepsilon$  and take a Taylor expansion around the data  $\mathbf{U}$

$$\begin{aligned} \mathbf{e}_\Theta - \mathbf{b}^\varepsilon &= (\mathbf{G} - \mathbf{G}^\star)\mathbf{w} + \text{vec}(\dot{\phi}\boldsymbol{\varepsilon}) \\ &= \left[ \mathbb{I}_d \otimes (\phi(\Theta(\mathbf{U}) - \Theta(\mathbf{U} - \boldsymbol{\varepsilon}))) \right] \mathbf{w} + \left[ \mathbb{I}_d \otimes \dot{\phi} \right] \text{vec}(\boldsymbol{\varepsilon}) \\ &= \mathbf{L}_\mathbf{w} \text{vec}(\boldsymbol{\varepsilon}) + \mathbf{h}(\mathbf{U}, \mathbf{w}, \boldsymbol{\varepsilon}) \end{aligned} \quad (11)$$

where  $\mathbf{h}(\mathbf{U}, \mathbf{w}, \boldsymbol{\varepsilon})$  is a vector-valued function of higher order terms in the measurement errors  $\boldsymbol{\varepsilon}$  (including the Hessian as well as higher order derivatives). Note that the  $\mathbf{h}$  function will generally produce a bias and higher-order dependencies for all system where  $\nabla^2\Theta \neq \mathbf{0}$ , but vanishes when  $\boldsymbol{\varepsilon} = \mathbf{0}$ .

The first order matrix in the expansion (11) is

$$\mathbf{L}_\mathbf{w} := [\text{mat}(\mathbf{w})^T \otimes \phi] \nabla \Theta \mathbf{K} + [\mathbb{I}_d \otimes \dot{\phi}],$$

where “mat” is the matricization operation and  $\mathbf{K}$  is the commutation matrix such that  $\mathbf{K}\text{vec}(\boldsymbol{\varepsilon}) = \text{vec}(\boldsymbol{\varepsilon}^T)$ . The matrix  $\nabla\Theta$  contains derivatives of the features

$$\nabla\Theta := \begin{bmatrix} \nabla f_1(\mathbf{U}_0) & & & \\ & \ddots & & \\ & & \nabla f_1(\mathbf{U}_M) & \\ & & \vdots & \\ \nabla f_J(\mathbf{U}_0) & & & \\ & \ddots & & \\ & & \nabla f_J(\mathbf{U}_M) & \end{bmatrix},$$

where

$$\nabla f_j(\mathbf{U}_m) = \left[ \frac{\partial}{\partial u_1} f_j(\mathbf{U}_m) \mid \cdots \mid \frac{\partial}{\partial u_d} f_j(\mathbf{U}_m) \right],$$

and  $\mathbf{U}_m \in \mathbb{R}^{1 \times d}$  is the row vector of data at  $t_m$ .

As mentioned above, we assume that all elements of  $\boldsymbol{\varepsilon}$  are i.i.d. Gaussian, i.e.,  $\mathcal{N}(0, \sigma^2)$  and thus to first order

$$\mathbf{r}(\mathbf{U}, \mathbf{w}) - (\mathbf{r}_0 + \mathbf{e}_{\text{int}}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{L}_{\mathbf{w}}(\mathbf{L}_{\mathbf{w}})^T). \quad (12)$$

In the case where  $\mathbf{w} = \mathbf{w}^*$  and the integration error is negligible, (12) simplifies to

$$\mathbf{G}\mathbf{w}^* - \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{L}_{\mathbf{w}^*}(\mathbf{L}_{\mathbf{w}^*})^T). \quad (13)$$

We note that the first order expansion in (11) performs particularly well when the underlying model equations are linear or a linearization of the equations is an accurate approximation. However, in the presence of strong model nonlinearities and large noise, this approximation is not substantially better than OLS. For instance, in the Hindmarsh-Rose example, with 128 data points and 10% noise (upper right of Fig. 8), the improvement over the OLS estimate of  $\mathbf{w}$  is less than 10%. Conversely, for this equation the issue can be resolved with higher resolution data (as illustrated by the higher resolution data performance also in Fig. 8).<sup>10</sup>

We note that in (13) (and in 12), the covariance is dependent upon the parameter vector  $\mathbf{w}$ . In the statistical inference literature, the Iteratively Reweighted Least Squares (IRLS) Jorgensen (2012) method offers a strategy to account for a parameter-dependent covariance by iterating between solving for  $\mathbf{w}$  and updating the covariance matrix  $\mathbf{C}$ . Furthermore, while the normality in (13) is approximate, the weighted least squares estimator has been shown to be consistent under fairly general conditions even without normality (Bollerslev and Wooldridge 1992). In Algorithm 2 we present the WENDy method, updating  $\mathbf{C}^{(n)}$  (at the  $n$ -th iteration step) in lines 7-8 and then the new parameters  $\mathbf{w}^{(n+1)}$  are computed in line 9 by weighted least squares.

<sup>10</sup> For a full description of the class of models for which more data always improves estimates, see Messenger and Bortz (2021b).

**Algorithm 2:** WENDy

```

input : Data  $\{\mathbf{U}\}$ , Feature Map  $\{\Theta, \nabla\Theta\}$ , Test Function Matrices  $\{\phi, \dot{\phi}\}$ , Stopping Criteria  $\{\text{SC}\}$ ,
        Covariance Relaxation Parameter  $\{\alpha\}$ , Variance Filter  $\{\mathbf{f}\}$ 
output: Parameter Estimate  $\{\hat{\mathbf{w}}, \hat{\mathbf{C}}, \hat{\sigma}, \mathbf{S}, \text{stdx}\}$ 

// Compute weak-form linear system
1  $\mathbf{G} \leftarrow [\mathbb{I}_d \otimes (\phi\Theta(\mathbf{U}))]$ 
2  $\mathbf{b} \leftarrow -\text{vec}(\dot{\phi}\mathbf{U})$ 

// Solve Ordinary Least Squares Problem
3  $\mathbf{w}^{(0)} \leftarrow (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \mathbf{b}$ 

// Solve Iteratively Reweighted Least Squares Problem
4  $n \leftarrow 0$ 
5  $\text{check} \leftarrow \text{true}$ 
6 while  $\text{check}$  is true do
7    $\mathbf{L}^{(n)} \leftarrow [\text{mat}(\mathbf{w}^{(n)})^T \otimes \phi] \nabla\Theta(\mathbf{U}) \mathbf{K} + [\mathbb{I}_d \otimes \dot{\phi}]$ 
8    $\mathbf{C}^{(n)} = (\mathbf{I} - \alpha) \mathbf{L}^{(n)} (\mathbf{L}^{(n)})^T + \alpha \mathbf{I}$ 
9    $\mathbf{w}^{(n+1)} \leftarrow (\mathbf{G}^T (\mathbf{C}^{(n)})^{-1} \mathbf{G})^{-1} \mathbf{G}^T (\mathbf{C}^{(n)})^{-1} \mathbf{b}$ 
10   $\text{check} \leftarrow \text{SC}(\mathbf{w}^{(n+1)}, \mathbf{w}^{(n)})$ 
11   $n \leftarrow n + 1$ 
12 end

// Return estimate and standard statistical quantities
13  $\hat{\mathbf{w}} \leftarrow \mathbf{w}^{(n)}$ 
14  $\hat{\mathbf{C}} \leftarrow \mathbf{C}^{(n)}$ 
15  $\hat{\sigma} \leftarrow (Md)^{-1/2} \|\mathbf{f} * \mathbf{U}\|_F$ 
16  $\mathbf{S} \leftarrow \hat{\sigma}^2 ((\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T) \hat{\mathbf{C}} (\mathbf{G} (\mathbf{G}^T \mathbf{G})^{-1})$ 
17  $\text{stdx} \leftarrow \sqrt{\text{diag}(\mathbf{S})}$ 

```

The IRLS step in line 9 requires inverting  $\mathbf{C}^{(n)}$ , which is done by computing its Cholesky factorization and then applying the inverse to  $\mathbf{G}$  and  $\mathbf{b}$ . Since this inversion may be unstable, we allow for possible regularization of  $\mathbf{C}^{(n)}$  in line 8 via a convex combination between the analytical first-order covariance  $\mathbf{L}^{(n)} (\mathbf{L}^{(n)})^T$  and the identity via the covariance relaxation parameter  $\alpha$ . This regularization allows the user to interpolate between the OLS solution ( $\alpha = 1$ ) and the unregularized IRLS solution ( $\alpha = 0$ ). In this way WENDy extends and encapsulates Algorithm 1. However, in the numerical examples below, we simply set  $\alpha = 10^{-10}$  throughout, as the aforementioned instability was not an issue. Lastly, any iterative scheme needs a stopping criteria and we will defer discussion of ours until Sect. 2.4.

The outputs of Algorithm 2 include the estimated parameters  $\hat{\mathbf{w}}$  as well as the covariance  $\hat{\mathbf{C}}$  of the response vector  $\mathbf{b}$  such that approximately

$$\mathbf{b} \sim \mathcal{N}(\mathbf{G}\hat{\mathbf{w}}, \hat{\sigma}^2 \hat{\mathbf{C}}). \quad (14)$$

A primary benefit of the WENDy methodology is that the parameter covariance matrix  $\mathbf{S}$  can be estimated from  $\hat{\mathbf{C}}$  using

$$\mathbf{S} := \hat{\sigma}^2 ((\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T) \hat{\mathbf{C}} (\mathbf{G} (\mathbf{G}^T \mathbf{G})^{-1}). \quad (15)$$

This yields the variances of individual components of  $\hat{\mathbf{w}}$  along  $\text{diag}(\mathbf{S})$  as well as the correlations between elements of  $\hat{\mathbf{w}}$  in the off-diagonals of  $\mathbf{S}$ . Here  $\hat{\sigma}^2$  is an estimate of the measurement variance  $\sigma^2$ , which we compute by convolving each compartment of the data  $\mathbf{U}$  with a high-order<sup>11</sup> filter  $\mathbf{f}$  and taking the Frobenius norm of the resulting convolved data matrix  $\mathbf{f} * \mathbf{U}$ . Throughout we set  $\mathbf{f}$  to be the centered finite difference weights of order 6 over 15 equally-spaced points (computed using Fornberg 1988), so that  $\mathbf{f}$  has order 5. The filter  $\mathbf{f}$  is then normalized to have unit 2-norm. This yields a high-accuracy approximation of  $\sigma^2$  for underlying data  $\mathbf{U}$  that is locally well-approximated by polynomials up to degree 5.

Once  $\mathbf{S}$  is obtained, for any given  $c \in (0, 1)$  one may compute a confidence interval  $[\hat{\mathbf{w}}_i - d_i(c), \hat{\mathbf{w}}_i + d_i(c)]$  around the learned parameter  $\hat{\mathbf{w}}_i$ , an interval which contains the ground truth parameter  $\mathbf{w}_i^*$  in  $100(1 - c)\%$  of trials under the assumption that  $\hat{\mathbf{w}}$  is normally distributed around  $\mathbf{w}^*$  with covariance matrix  $\mathbf{S}$ . For  $0 \leq c \leq 1$ , the bound  $d_i(c)$  is defined by

$$d_i(c) = F_{\mathbf{S}_{ii}}^{-1}(1 - c/2) \quad (16)$$

where  $F_{\mathbf{S}_{ii}}(x) = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{x}{\sqrt{2\mathbf{S}_{ii}}} \right) \right]$  is the CDF of a normal distribution with mean zero and variance  $\mathbf{S}_{ii}$ .

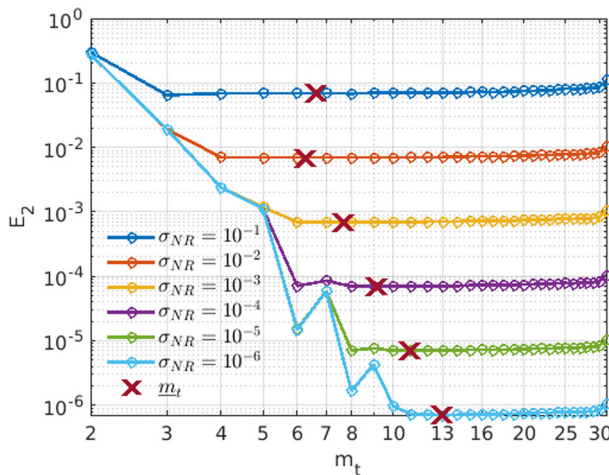
Note the above provides only individual parameters' confidence intervals. In general, if multivariate confidence regions are of interest, they can be obtained using the  $F$  distribution, or Hotelling's  $T$ -squared distribution. The latter is able to account for the uncertainty in the estimated variance-covariance matrix  $\mathbf{S}$  simultaneously with the joint uncertainty in the vector  $\hat{\mathbf{w}}$ .

### 2.3 Choice of Test Functions

When using WENDy for parameter estimation, a valid question concerns the choice of test function. This is particularly challenging in the sparse data regime, where integration errors can easily affect parameter estimates. In Messenger and Bortz (2021b) we reported that using higher order polynomials as test functions yielded more accuracy (up to machine precision). Inspired by this result and to render moot the question of what order polynomial is needed, we have developed a 2-step process for offline computation of highly efficient test functions, given a timegrid  $\mathbf{t}$ .

We note that in (9) when there is no noise, the only remaining term in the residual is the integration error  $\mathbf{e}_{\text{int}}$ . We can derive an estimator that can be computed using the noisy data  $\mathbf{U}$  and used to detect a minimal radius  $\underline{m}_t$  such that  $m_t > \underline{m}_t$  leads to negligible integration error compared to the errors introduced by random noise. Inspired by wavelet decompositions, we next row-concatenate convolution matrices of test functions at different radii  $\mathbf{m}_t := (2^\ell \underline{m}_t; \ell = \{0, \dots, \bar{\ell}\})$ . An SVD of this tall matrix yields an orthonormal test function matrix  $\phi$ , which maximally extracts information across different scales. We note that in the later examples we have  $\bar{\ell} = 3$ ,

<sup>11</sup> The order of a filter is defined as the number of moments that the filter leaves zero (other than the zero-th moment). For more mathematical details see (Messenger and Bortz 2022a) Appendix F.



**Fig. 1** Coefficient error  $E_2 = \|\mathbf{w}^* - \hat{\mathbf{w}}\|_2 / \|\mathbf{w}^*\|_2$  of WENDy applied to the Logistic Growth model vs test function radius  $m_t$  for noise levels  $\sigma_{NR} \in \{10^{-6}, \dots, 10^{-1}\}$ . For large enough radius, errors are dominated by noise and integration error is negligible. The minimum radius  $\underline{m}_t$  computed as in Sect. 2.3.1 finds this noise-dominated region, which varies depending on  $\sigma_{NR}$

which in many cases leads to a largest test function support covering half of the time domain.

To begin, we consider a  $C^\infty$  bump function

$$\psi(t; a) = C \exp\left(-\frac{\eta}{[1 - (t/a)^2]_+}\right), \quad (17)$$

where the constant  $C$  enforces that  $\|\psi\|_2 = 1$ ,  $\eta$  is a shape parameter, and  $[\cdot]_+ := \max(\cdot, 0)$ , so that  $\psi(t; a)$  is supported only on  $[-a, a]$  where

$$a = m_t \Delta t. \quad (18)$$

With the  $\psi$  in (17) we have discovered that the accuracy of the parameter estimates is relatively insensitive to a wide range of  $\eta$  values. Therefore, based on empirical investigation we arbitrarily choose  $\eta = 9$  in all examples and defer more extensive analysis to future work. In the rest of this section, we will describe the computation of  $\underline{m}_t$  and how to use  $\psi$  to construct  $\phi$  and  $\dot{\phi}$ .

### 2.3.1 Minimum Radius Selection

In (9), the residual is decomposed into several terms. Notably the  $\mathbf{e}_{\text{int}}$  term can be interpreted as the error in the residual for a specified test function at the true solution  $\mathbf{u}^*$ . Below we show how to reduce this component of the residual. Figure 1 illustrates for the Logistic Growth model how the relative error changes as a function of test function radius  $m_t$  (for different noise levels). As the radius increases, the error becomes dominated by the measurement noise. To establish a lower bound  $\underline{m}_t$  on the

test function radius  $m_t$ , we create an estimate for the integration error which works for any of the  $d$  variables in a model. To promote clarity, we will let  $u$  be any of the  $d$  variables for the remainder of this section. However, it is important to note the the final  $\widehat{\mathbf{e}}_{\text{rms}}$  sums over all  $d$  variables.

We now consider the  $k$ -th element of  $\mathbf{e}_{\text{int}}$

$$\begin{aligned}\mathbf{e}_{\text{int}}(u^*, \phi_k, M)_k &= (\mathbf{G}^* \mathbf{w}^* - \mathbf{b}^*)_k = \sum_{m=0}^{M-1} (\phi_k(t_m) \dot{\mathbf{u}}_m^* + \dot{\phi}_k(t_m) \mathbf{u}_m^*) \\ \Delta t &= \frac{T}{M} \sum_{m=0}^{M-1} \frac{d}{dt} (\phi_k(t_m) \mathbf{u}_m^*),\end{aligned}$$

where  $\Delta t = T/M$  for a uniform timegrid  $\mathbf{t} = (0, \Delta t, 2\Delta t, \dots, M\Delta t)$  with overall length  $T$ . We also note that the biggest benefit of this approach is that  $\mathbf{e}_{\text{int}}$  does not explicitly depend upon  $\mathbf{w}^*$ .

By expanding  $\frac{d}{dt}(\phi_k(t)u^*(t))$  into its Fourier series<sup>12</sup> we then have

$$\begin{aligned}\mathbf{e}_{\text{int}}(u^*, \phi_k, M) &= \frac{T}{M\sqrt{T}} \sum_{n \in \mathbb{Z}} \mathcal{F}_n \left[ \frac{d}{dt}(\phi_k(t)u^*(t)) \right] \left( \sum_{m=0}^{M-1} e^{2\pi i n m / M} \right) \\ &= \frac{2\pi i}{\sqrt{T}} \sum_{n \in \mathbb{Z}} n M \mathcal{F}_{nM}[\phi_k u^*],\end{aligned}\quad (19)$$

so that the integration error is entirely represented by aliased modes  $\{M, 2M, \dots\}$  of  $\phi_k u^*$ . For  $a > 0$  as defined in (18), if  $[t_k - a, t_k + a] \subset [0, T]$  and  $T/2 > a > 1$ , we have the relation

$$\begin{aligned}\mathcal{F}_n[\phi_k(\cdot; a)] &= \frac{a}{\sqrt{T}} \int_0^{T/a} \phi(s; 1) e^{-2\pi i n a / T} ds \\ &= \frac{a}{\sqrt{T}} \int_0^T \phi(s; 1) e^{-2\pi i n a / T} ds \\ &= a \mathcal{F}_{na}[\phi_k(\cdot; 1)]\end{aligned}$$

where the first equality comes from the change of variables  $s = t/a$  and the second from extending the domain of integration to  $[0, T]$  using compact support of  $\phi$ . This suggests that increasing  $m_t$  corresponds to higher-order Fourier coefficients of  $\phi_k(\cdot; 1)$  entering the error formula (19), which shows, using (19), that increasing  $a$  (eventually) lowers the integration error. For small  $m_t$ , this leads to the  $\mathbf{e}_{\text{int}}$  term being dominated by the numerical integration approximation error, while for large  $m_t$ , the noise-related effects are dominant.

<sup>12</sup> For  $f \in L^2([0, T])$  (i.e.  $\int_0^T f^2(t)dt < \infty$ ) we define the  $\xi$ th Fourier coefficient of  $f$  for  $\xi \in \mathbb{R}$  as  $\mathcal{F}_\xi[f] := \frac{1}{\sqrt{T}} \int_0^T f(t) e^{-2\pi i \xi / T} dt$ , and  $f$  is equal to its Fourier series,  $f(t) = \frac{1}{\sqrt{T}} \sum_{n \in \mathbb{Z}} \mathcal{F}_n[f] e^{2\pi i n / T}$ . We refer to  $\mathcal{F}_n[f]$  as the  $n$ th Fourier mode of  $f$  when  $n$  is an integer.

We now derive a surrogate approximation of  $\mathbf{e}_{\text{int}}$  using the noisy data  $\mathbf{U}$  to estimate this transition from integration error-dominated to noise error-dominated residuals. From the noisy data  $\mathbf{U}$  on timegrid  $\mathbf{t} \in \mathbb{R}^M$ , we wish to compute  $\mathbf{e}_{\text{int}}(u^*, \phi_k, M)$  by substituting  $\mathbf{U}$  for  $u^*$  and using the discrete Fourier transform (DFT), however the highest mode<sup>13</sup> we have access to is  $\widehat{\mathcal{F}}_{\pm M/2}[\phi \mathbf{U}]$ . On the other hand, we *are* able to approximate  $\mathbf{e}_{\text{int}}(u^*, \phi_k, \lfloor M/s \rfloor)$  from  $\mathbf{U}$ , that is, the integration error over a *coarsened* timegrid  $(0, \widetilde{\Delta t}, 2\widetilde{\Delta t}, \dots, \lfloor M/s \rfloor \widetilde{\Delta t})$ , where  $\widetilde{\Delta t} = T/\lfloor M/s \rfloor$  and  $s > 2$  is a chosen coarsening factor. By introducing the truncated error formula

$$\widehat{\mathbf{e}}_{\text{int}}(u^*, \phi_k, \lfloor M/s \rfloor, s) := \frac{2\pi i}{\sqrt{T}} \sum_{n=-\lfloor s/2 \rfloor}^{\lfloor s/2 \rfloor} n \lfloor M/s \rfloor \mathcal{F}_{n \lfloor M/s \rfloor}[\phi_k u^*],$$

we have that

$$\widehat{\mathbf{e}}_{\text{int}}(u^*, \phi_k, \lfloor M/s \rfloor, s) \approx \mathbf{e}_{\text{int}}(u^*, \phi_k, \lfloor M/s \rfloor),$$

and  $\widehat{\mathbf{e}}_{\text{int}}$  can be directly evaluated at  $\mathbf{U}$  using the DFT. In particular, with  $2 < s < 4$ , we get

$$\begin{aligned} \widehat{\mathbf{e}}_{\text{int}}(\mathbf{U}, \phi_k, \lfloor M/s \rfloor, s) &= \frac{2\pi i \lfloor M/s \rfloor}{\sqrt{T}} (\widehat{\mathcal{F}}_{\lfloor M/s \rfloor}[\phi_k \mathbf{U}] - \widehat{\mathcal{F}}_{-\lfloor M/s \rfloor}[\phi_k \mathbf{U}]) \\ &= -\frac{4\pi \lfloor M/s \rfloor}{\sqrt{T}} \text{Im}\{\widehat{\mathcal{F}}_{\lfloor M/s \rfloor}[\phi_k \mathbf{U}]\} \end{aligned}$$

where  $\text{Im}\{z\}$  denotes the imaginary portion of  $z \in \mathbb{C}$ , so that only a single Fourier mode needs computation. In most practical cases of interest, this leads to (see Fig. 2)

$$\mathbf{e}_{\text{int}}(u^*, \phi_k, M) \leq \widehat{\mathbf{e}}_{\text{int}}(\mathbf{U}, \phi_k, \lfloor M/s \rfloor, s) \leq \mathbf{e}_{\text{int}}(u^*, \phi_k, \lfloor M/s \rfloor) \quad (20)$$

so that ensuring  $\widehat{\mathbf{e}}_{\text{int}}(\mathbf{U}, \phi_k, \lfloor M/s \rfloor, s)$  is below some tolerance  $\tau$  leads also to  $\mathbf{e}_{\text{int}}(u, \phi_k, M) < \tau$ .

Statistically, under our additive noise model we have that  $\widehat{\mathbf{e}}_{\text{int}}(\mathbf{U}, \phi_k, \lfloor M/s \rfloor, s)$  is an unbiased estimator of  $\widehat{\mathbf{e}}_{\text{int}}(u^*, \phi_k, \lfloor M/s \rfloor, s)$ , i.e.,

$$\begin{aligned} \mathbb{E}[\widehat{\mathbf{e}}_{\text{int}}(\mathbf{U}, \phi_k, \lfloor M/s \rfloor, s)] &= \mathbb{E}[-(4\pi \lfloor M/s \rfloor / \sqrt{T}) \text{Im}\{\widehat{\mathcal{F}}_{\lfloor M/s \rfloor}[\phi_k (u^* + \boldsymbol{\varepsilon})]\}] \\ &= \mathbb{E}[\widehat{\mathbf{e}}_{\text{int}}(u^*, \phi_k, \lfloor M/s \rfloor, s)], \end{aligned}$$

<sup>13</sup> We define the  $n$ th discrete Fourier mode of a function  $f$  over a periodic grid  $(m\Delta t)_{m=0}^M$  by  $\widehat{\mathcal{F}}_n[f] := \frac{\Delta t}{\sqrt{M\Delta t}} \sum_{m=0}^{M-1} f(m\Delta t) e^{-2\pi i n m / M}$ .



where  $\mathbb{E}$  denotes expectation. The variance satisfies, for  $2 < s < 4$ ,

$$\begin{aligned} \text{Var}[\hat{\mathbf{e}}_{\text{int}}(\mathbf{U}, \phi_k, \lfloor M/s \rfloor, s)] &:= \sigma^2 \left( \frac{4\pi \lfloor M/s \rfloor}{M} \right)^2 \\ \sum_{j=1}^{M-1} \phi_k^2(j \Delta t) \sin^2(2\pi \lfloor M/s \rfloor j/M) &\leq \sigma^2 \left( \frac{4\pi \lfloor M/s \rfloor}{M} \right)^2 \end{aligned}$$

where  $\sigma^2 = \text{Var}[\epsilon]$ . The upper bound follows from  $\|\phi_k\|_2 = 1$ , and shows that the variance is not sensitive to the radius of the test function  $\phi_k$ .

We pick a minimum radius  $\underline{m}_t$  as a changepoint of  $\log(\hat{\mathbf{e}}_{\text{rms}})$ , where  $\hat{\mathbf{e}}_{\text{rms}}$  is the root-mean-squared integration error over test functions placed along the timeseries,

$$\hat{\mathbf{e}}_{\text{rms}}(m_t) := K^{-1} \sum_{k=1}^K \sum_{i=1}^d \hat{\mathbf{e}}_{\text{int}}(\mathbf{U}^{(i)}, \phi_k(\cdot; m_t), \lfloor M/s \rfloor, s)^2, \quad (21)$$

where  $\mathbf{U}^{(i)}$  is the  $i$ th variable in the system. Figure 2 depicts  $\hat{\mathbf{e}}_{\text{rms}}$  as a function of support radius  $m_t$ . As can be seen, since the variance of  $\hat{\mathbf{e}}_{\text{int}}$  is insensitive to the radius  $m_t$ , the estimator is approximately flat over the region with negligible integration error, a perfect setting for changepoint detection. Crucially, Figure 2 demonstrates that, in practice, the minimum radius  $\underline{m}_t$  lies to the right of the changepoint of the coefficient errors

$$E_2(\hat{\mathbf{w}}) := \|\hat{\mathbf{w}} - \mathbf{w}^*\|_2^2 / \|\mathbf{w}^*\|_2^2,$$

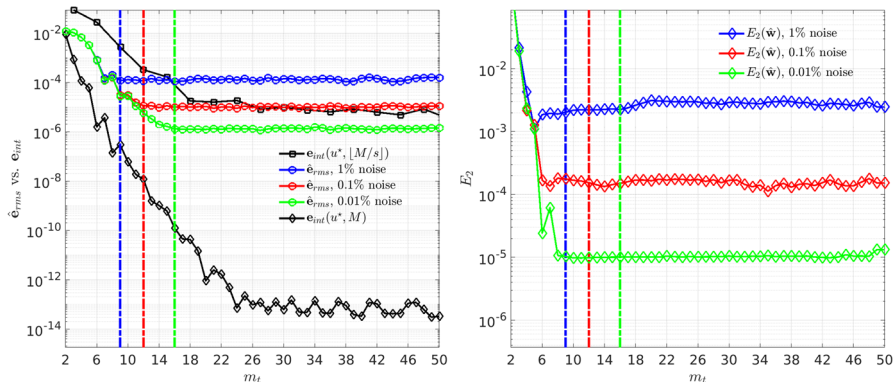
as a function of  $m_t$ . Lastly, note that the red  $\times$  in Figure 1 depicts the identified  $\underline{m}_t$  for the Logistic Growth model.

### 2.3.2 Orthonormal Test Functions

Having computed the minimal radius  $\underline{m}_t$ , we then construct the test function matrices  $(\phi, \dot{\phi})$  by orthonormalizing and truncating a concatenation of test function matrices with  $\mathbf{m}_t := \underline{m}_t \times (1, 2, 4, 8)$ . Letting  $\Psi_\ell$  be the convolution matrix for  $\psi(\cdot; 2^\ell \underline{m}_t \Delta t)$ , we compute the SVD of

$$\Psi := \begin{bmatrix} \Psi_0 \\ \Psi_1 \\ \Psi_2 \\ \Psi_3 \end{bmatrix} = \mathbf{Q} \mathbf{\Sigma} \mathbf{V}^T.$$

The right singular vectors  $\mathbf{V}$  then form an orthonormal basis for the set of test functions forming the rows of  $\Psi$ . Letting  $r$  be the rank of  $\Psi$ , we then truncate the SVD to rank  $K$ , where  $K$  is selected as the changepoint in the cumulative sum of the singular values



**Fig. 2** Visualization of the minimum radius selection using single realizations of Fitzhugh-Nagumo data with 512 timepoints at three different noise levels. Dashed lines indicate the minimum radius  $\underline{m}_t$ . Left: we see that inequality (20) holds empirically for small radii  $m_t$ . Right: coefficient error  $E_2$  as a function of  $m_t$  is plotted, showing that for each noise level the identified radius  $m_t$  using  $\hat{e}_{rms}$  lies to right of the dip in  $E_2$ , as random errors begin to dominate integration errors. In particular, for low levels of noise,  $\underline{m}_t$  increases to ensure high accuracy integration

$(\Sigma_{ii})_{i=1}^r$ . We then let

$$\phi = (\mathbf{V}^{(K)})^T$$

be the test function basis where  $\mathbf{V}^{(K)}$  indicates the first  $K$  modes of  $\mathbf{V}$ . Unlike our previous implementations, the derivative matrix  $\dot{\phi}$  must now be computed numerically, however given the compact support and smoothness of the reference test functions  $\psi(\cdot; 2^\ell \underline{m}_t \Delta t)$ , this can be done very accurately with Fourier differentiation. Hence, we let

$$\dot{\phi} = \mathcal{F}^{-1} \text{diag}(i\mathbf{k}) \mathcal{F} \phi$$

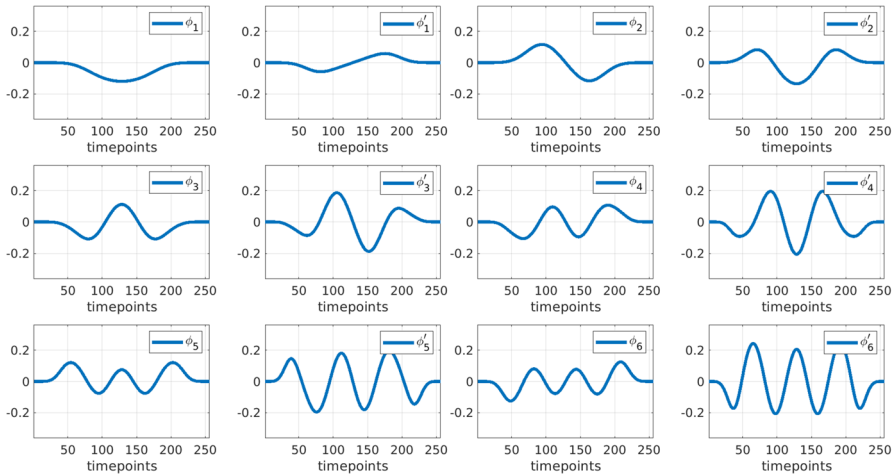
where  $\mathcal{F}$  is the discrete Fourier transform and  $\mathbf{k}$  are the requisite wavenumbers. Figure 3 displays the first six orthonormal test functions along with their derivatives obtained from this process applied to Hindmarsh-Rose data.

## 2.4 Stopping Criteria

Having formed the test function matrices  $\{\phi, \dot{\phi}\}$ , the remaining unspecified process in Algorithm 2 is the stopping criteria SC. The iteration can stop in one of three ways: (1) the iterates reach a fixed point, (2) the number of iterates exceeds a specified limit, or (3) the residuals

$$\mathbf{r}^{(n+1)} := (\mathbf{C}^{(n)})^{-1/2} (\mathbf{G}\mathbf{w}^{(n+1)} - \mathbf{b})$$

are no longer approximately normally distributed. (1) and (2) are straightforward limitations of any iterative algorithm while (3) results from the fact that our weighted



**Fig. 3** First six orthonormal test functions obtained from Hindmarsh-Rose data with 2% noise and 256 timepoints using the process outlined in Sect. 2.3.2

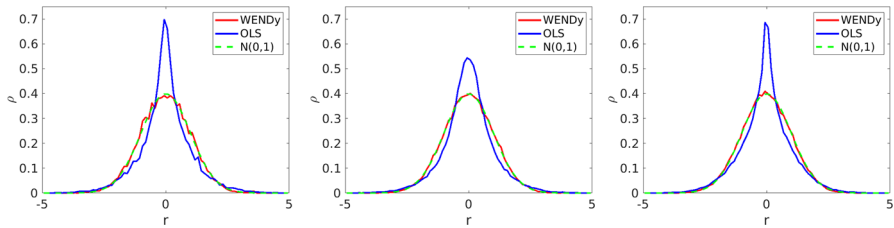
least-squares framework is only approximate. In ideal scenarios where the discrepancy terms  $\mathbf{e}_{\text{int}}$  and  $\mathbf{h}(\mathbf{u}^*, \mathbf{w}^*; \boldsymbol{\varepsilon})$  are negligible, Eq. (12) implies that

$$(\mathbf{C}^*)^{-1}(\mathbf{G}\mathbf{w}^* - \mathbf{b}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

where  $\mathbf{C}^* = \mathbf{L}^*(\mathbf{L}^*)^T$  is the covariance computed from  $\mathbf{w}^*$ . Hence we expect  $\mathbf{r}^{(n)}$  to agree with a normal distribution more strongly as  $n$  increases. If the discrepancy terms are non-negligible, it is possible that the reweighting procedure will not result in an increasingly normal  $\mathbf{r}^{(n)}$ , and iterates  $\mathbf{w}^{(n)}$  may become worse approximations of  $\mathbf{w}^*$ . A simple way to detect this is with the Shapiro-Wilk (S-W) test for normality (Shapiro and Wilk 1965), which produces an approximate  $p$ -value under the null hypothesis that the given sample is i.i.d. normally distributed. However, the first few iterations are also not expected to yield i.i.d. normal residuals (see Figure 4), so we only check the S-W test after a fixed number of iterations  $n_0$ . Letting  $\text{SW}^{(n)} := \text{SW}(\mathbf{r}^{(n)})$  denote the  $p$ -value of the S-W test at iteration  $n > n_0$ , and setting  $\text{SW}^{(n_0)} = 1$ , we specify the stopping criteria as:

$$\begin{aligned} \text{SC}(\mathbf{w}^{(n+1)}, \mathbf{w}^{(n)}) = \{ \|\mathbf{w}^{(n+1)} - \mathbf{w}^{(n)}\|_2 / \|\mathbf{w}^{(n)}\|_2 > \tau_{\text{FP}} \} \text{ and } \{ n < \text{max\_its} \} \\ \text{and } \{ \text{SW}^{(\max\{n, n_0\})} > \tau_{\text{SW}} \}. \end{aligned} \quad (22)$$

We set the fixed-point tolerance to  $\tau_{\text{FP}} = 10^{-6}$ , the S-W tolerance and starting point to  $\tau_{\text{SW}} = 10^{-4}$  and  $n_0 = 10$ , and  $\text{max\_its} = 100$ .



**Fig. 4** Histograms of the OLS residual in blue ( $\mathbf{r}(\mathbf{U}, \mathbf{w}^*)$ , Eq. 8) versus the WENDy residual in red (given by  $\mathbf{C}(\mathbf{w}^*)^{-1/2}\mathbf{r}(\mathbf{U}, \mathbf{w}^*)$ , see Eq. (14)) both evaluated at the true solution  $\mathbf{w}^*$ , overlaying the probability density function of a standard normal  $\mathcal{N}(0, 1)$  in green. Each curve approximates the probability density  $\rho(r)$  of finding a residual entry near  $r$  under the given statistical model. Left to right: Logistic Growth, Lotka-Volterra, and Fitzhugh-Nagumo, each with 256 timepoints and 20% noise. Curves are averaged over 500 independent trials with each histogram scaled by its empirical standard deviation. In each case, the WENDy residual agrees well with a standard normal, while the OLS residual exhibits distinctly non-Gaussian features, indicative that OLS is the wrong statistical regression model (Color figure online)

## 2.5 Comments on the Convergence of the WENDy Estimate

In §2.1, Conditions 1-4 describe the criteria needed for the OLS problem to have a unique solution. However, the EiV- / IRLS-based WENDy method in Algorithm 2 is iterative. To ensure that WENDy converges to a unique solution would mean proving that the algorithm is a contraction map converging to a fixed point when the initial estimate (i.e., the OLS estimate) is close enough to the true solution  $\mathbf{w}^*$ . This is not straightforward, given that the covariance is updated at every step and we thus leave as a topic for future work.

## 3 Illustrating Examples

Here we demonstrate the effectiveness of WENDy applied to five ordinary differential equations canonical to biology and biochemical modeling (see Table 1 for the specific equations and parameters used). As demonstrated in the works mentioned in Sect. 1, it is known that the weak or integral formulations are advantageous, with previous works mostly advocating for a two step process involving (1) pre-smoothing the data before (2) solving for parameters using ordinary least squares. The WENDy approach does not involve smoothing the data, and instead leverages the covariance structure introduced by the weak form to iteratively reduce errors in the ordinary least squares (OLS) weak-form estimation. Utilizing the covariance structure in this way not only reduces error, but reveals parameter uncertainties as demonstrated in Sect. 3.3.

We compare the WENDy solution to the weak-form ordinary least squares solution (described in Sect. 2 and denoted simply by OLS in this section) to forward solver-based nonlinear least squares (FSNLS). Comparison to OLS is important due to the growing use of weak formulations in joint equation learning/parameter estimation tasks, but often without smoothing or further variance reduction steps (Messenger and Bortz 2021a; Fasel et al. 2021; Nicolaou et al. 2023; Bertsimas and Gurnee 2023). In most cases WENDy reduces the OLS error by 60%–90% (see the bar plots in Figs.

Table 1 Specifications of ODE examples

Name	ODE	Parameters
Logistic growth	$\dot{u} = w_1 u + w_2 u^2$	$T = 10, u(0) = 0.01,$ $\ \text{vec}(\mathbf{U}^*)\ _{\text{rms}} = 0.66,$ $\mathbf{w}^* = (1, -1)$
Lotka-volterra	$\begin{cases} \dot{u}_1 = w_1 u_1 + w_2 u_1 u_2 \\ \dot{u}_2 = w_3 u_2 + w_4 u_1 u_2 \end{cases}$	$T = 5, u(0) = (1, 1),$ $\ \text{vec}(\mathbf{U}^*)\ _{\text{rms}} = 6.8,$ $\mathbf{w}^* = (3, -1, -6, 1)$
Fitzhugh-nagumo	$\begin{cases} \dot{u}_1 = w_1 u_1 + w_2 u_1^3 + w_3 u_2 \\ \dot{u}_2 = w_4 u_1 + w_5(1) + w_6 u_2 \end{cases}$	$T = 25, u(0) = (0, 0.1),$ $\ \text{vec}(\mathbf{U}^*)\ _{\text{rms}} = 0.68, \mathbf{w}^* =$ $(3, -3, 3, -1/3, 17/150, 1/15)$ $T = 10,$ $u(0) = (-1.31, -7.6, -0.2),$ $\ \text{vec}(\mathbf{U}^*)\ _{\text{rms}} = 2.8, \mathbf{w}^* =$ $(10, -10, 30, -10, 10, -50, -10,$ $0.04, 0.0319, -0.01)$
Hindmarsh-rose	$\begin{cases} \dot{u}_1 = w_1 u_2 + w_2 u_1^3 + w_3 u_1^2 + w_4 u_3 \\ \dot{u}_2 = w_5(1) + w_6 u_1^2 + w_7 u_2 \\ \dot{u}_3 = w_8 u_1 + w_9(1) + w_{10} u_3 \end{cases}$	
Protein transduction benchmark (PTB)	$\begin{cases} \dot{u}_1 = w_1 u_1 + w_2 u_1 u_3 + w_3 u_4 \\ \dot{u}_2 = w_4 u_1 \\ \dot{u}_3 = w_5 u_1 u_3 + w_6 u_4 + w_7 \frac{u_5}{0.3+u_5} \\ \dot{u}_4 = w_8 u_1 u_3 + w_9 u_4 \\ \dot{u}_5 = w_{10} u_4 + w_{11} \frac{u_5}{0.3+u_5} \end{cases}$	$T = 25, u(0) = (1, 0, 1, 0, 1),$ $\ \text{vec}(\mathbf{U}^*)\ _{\text{rms}} = 0.81, \mathbf{w}^* =$ $(-0.07, -0.6, 0.35, 0.07, -0.6, 0.05,$ $0.17, 0.6, -0.35, 0.3, -0.017)$

Note that  $\|\text{vec}(\mathbf{U}^*)\|_{\text{rms}}$  is included for reference in order to compute the noise variance using  $\sigma = \sigma_{N/R} / \|\text{vec}(\mathbf{U}^*)\|_{\text{rms}}$

5, 6, 7, 8 and 9). When compared to FSNLS, WENDy provides a more efficient and accurate solution in typical use cases, however in the regime of highly sparse data and large noise, FSNLS provides an improvement in accuracy at a higher computational cost. Furthermore, we demonstrate that FSNLS may be improved by using the WENDy output as an initial guess. We aim to explore further benefits of combining forward solver-based approaches with solver-free weak-form approaches in a future work. Code to generate all examples is available at <https://github.com/MathBioCU/WENDy>.

### 3.1 Numerical Methods and Performance Metrics

In all cases below, we solve for approximate weights  $\hat{\mathbf{w}}$  using Algorithm 2 over 100 independent trials of additive Gaussian noise with standard deviation  $\sigma = \sigma_{NR} \|\text{vec}(\mathbf{U}^*)\|_{\text{rms}}$  for a range of noise ratios  $\sigma_{NR}$ . This specification of the variance implies that

$$\sigma_{NR} \approx \frac{\|\text{vec}(\mathbf{U}^* - \mathbf{U})\|_{\text{rms}}}{\|\text{vec}(\mathbf{U})\|_{\text{rms}}},$$

so that  $\sigma_{NR}$  can be interpreted as the relative error between the true and noisy data. Results from all trials are aggregated by computing the mean and median. Computations of Algorithm 2 are performed in MATLAB on a laptop with 40GB of RAM and an 8-core AMD Ryzen 7 pro 4750u processor. Computations of FSNLS are also performed in MATLAB but were run on the University of Colorado Boulder's Blanca Condo Cluster in a trivially parallel manner over a homogeneous CPU set each with Intel Xeon Gold 6130 processors and 24GB RAM. Due to the comparable speed of the two processors (1.7 GHz for AMD Ryzen 7, 2.1 GHz for Intel Xeon Gold) and the fact that each task required less than 5 GB working memory (well below the maximum allowable), we believe the walltime comparisons between WENDy and FSNLS below are fair.

As well as  $\sigma_{NR}$ , we vary the stepsize  $\Delta t$  (keeping the final time  $T$  fixed for each example), to demonstrate large and small sample behavior. For each example, a high-fidelity solution is obtained on a fine grid (512 timepoints for Logistic Growth, 1024 for all other examples), which is then subsampled by factors of 2 to obtain coarser datasets.

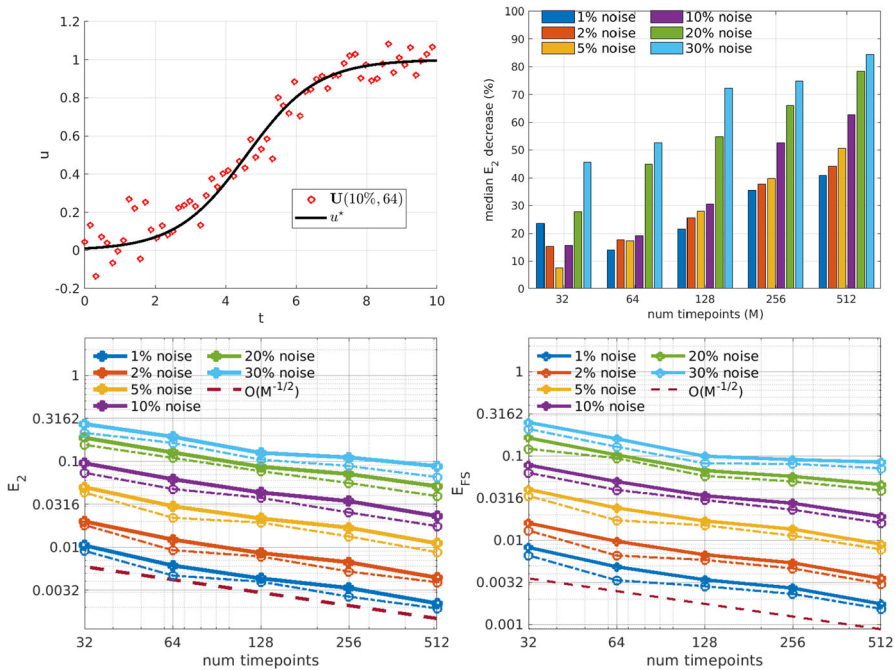
To evaluate the performance of WENDy, we record the relative coefficient error

$$E_2 := \frac{\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2}{\|\mathbf{w}^*\|_2} \quad (23)$$

as well as the forward simulation error

$$E_{\text{FS}} := \frac{\|\text{vec}(\mathbf{U}^* - \hat{\mathbf{U}})\|_2}{\|\text{vec}(\mathbf{U}^*)\|_2}. \quad (24)$$

The data  $\hat{\mathbf{U}}$  is obtained by simulating forward the model using the learned coefficients  $\hat{\mathbf{w}}$  from the exact initial conditions  $u(0)$  using the same  $\Delta t$  as the data. The RK45



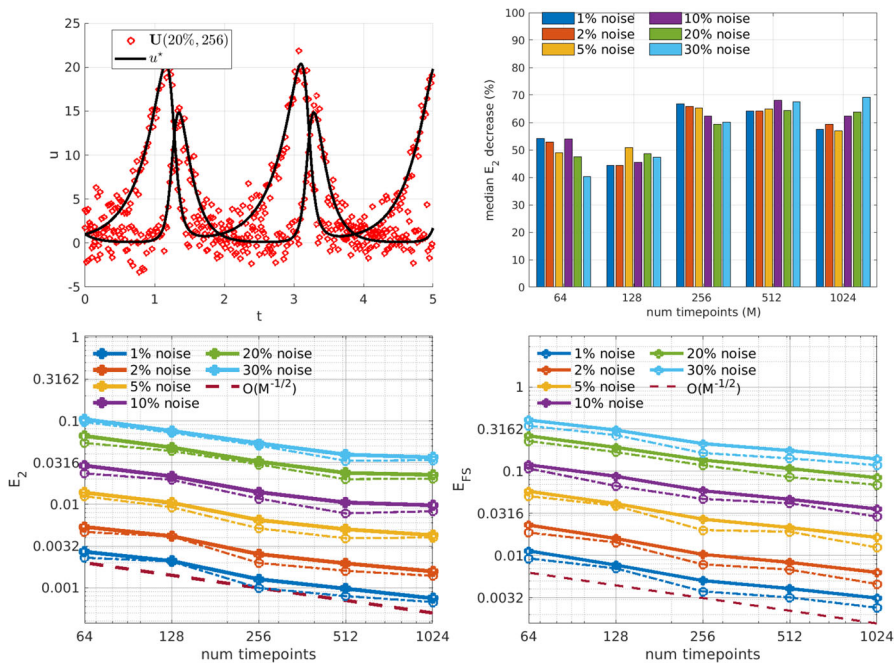
**Fig. 5** Logistic growth: Estimation of parameters in the Logistic Growth model. Top left: true solution with example noise realization. Top right: median percentage drop in  $E_2$  from the OLS solution to the WENDy output (e.g. at 30% noise and 512 timepoints WENDy results in a 85% reduction in error). Bottom left and right panels display parameter errors  $E_2$  and forward simulation error  $E_{FS}$ , respectively. Solid lines show the mean error and dash-dot lines show the median error. The dashed maroon line depicts an  $\mathcal{O}(M^{-1/2})$  curve

algorithm is used for all forward simulations (unless otherwise specified) with relative and absolute tolerances of  $10^{-12}$ . Comparison with OLS solutions is displayed in bar graphs which give the drop in error from the OLS solution to the WENDy solution as a percentage of the error in the OLS solution (Figs. 5, 6, 7, 8, 9).

## 3.2 Summary of Results

### 3.2.1 Logistic Growth

The logistic growth model is the simplest nonlinear model for population growth, yet the  $u^2$  nonlinearity generates a bias that affects the OLS solution more strongly as noise increases. Figure 5 (top right) indicates that when  $M \geq 256$  WENDy decreases the error by 50%-85% from the OLS solution for noise level is 10% or higher. WENDy also leads to a robust fit for smaller  $M$ , providing coefficient errors  $E_2$  and forward simulation errors  $E_{FS}$  that are both less than 6% for data with only 64 points and 10% noise (Fig. 5 (top left) displays an example dataset at this resolution).



**Fig. 6** Lotka-volterra: Estimation of parameters in the Lotka-Volterra model (for plot details see Figure 5 caption)

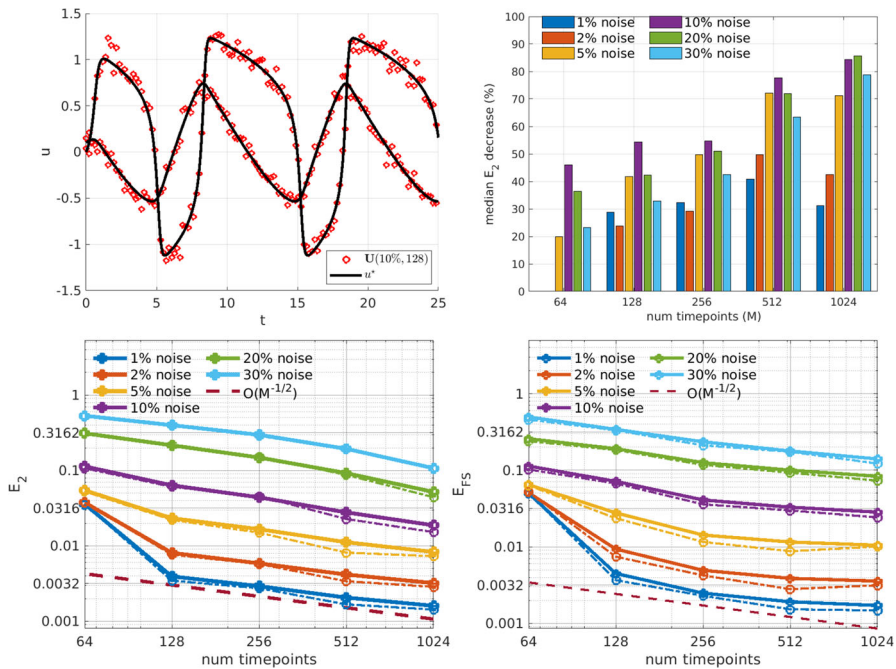
### 3.2.2 Lotka-Volterra

The Lotka-Volterra model is a system of equations designed to capture predator-prey dynamics (Lotka 1978). Each term in the model is unbiased when evaluated at noisy data (under the i.i.d. assumption), so that the first-order residual expansion utilized in WENDy is highly accurate. The bottom left plot in Fig. 6 shows even with 30% noise and only 64 timepoints, the coefficient error is still less than 10%. WENDy reduces the error by 40%–70% on average from the OLS (top right panel).

### 3.2.3 Fitzhugh-Nagumo

The Fitzhugh-Nagumo equations are a simplified model for an excitable neuron (FitzHugh 1961). The equations contain six fundamental terms with coefficients to be identified. The cubic nonlinearity implies that the first-order covariance expansion in WENDy becomes inaccurate at high levels of noise. Nevertheless, Fig. 7 (lower plots) shows that WENDy produces on average 6% coefficient errors at 10% noise with only 128 timepoints, and only 7% forward simulation errors (see upper left plot for an example dataset at this resolution). In many cases WENDy reduces the error by over 50% from the OLS solution, with 80% reductions for high noise and  $M = 1024$  timepoints (top right panel). For sparse data (e.g. 64 timepoints), numerical integration



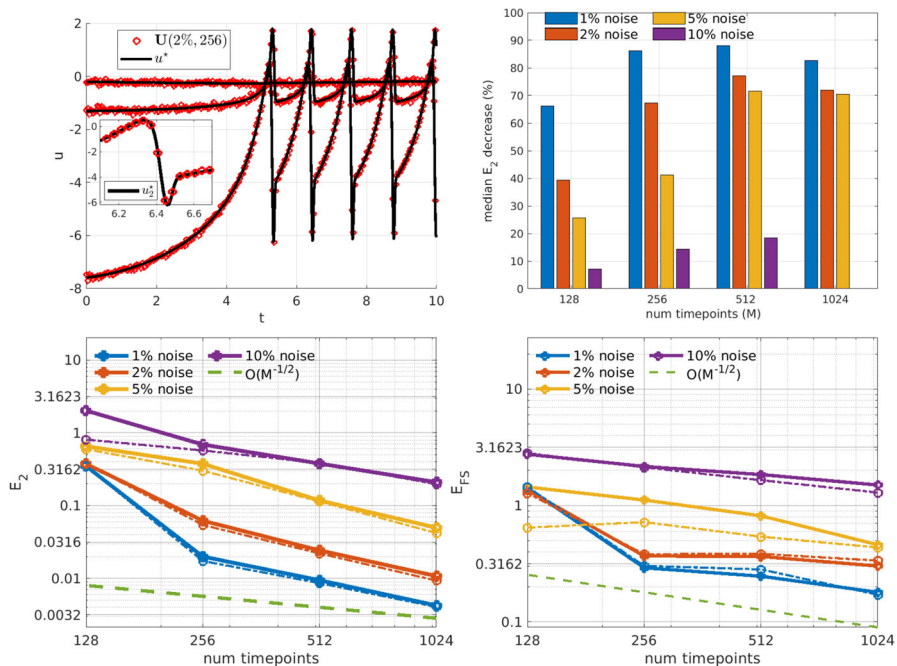


**Fig. 7** FitzHugh-nagumo: Estimation of parameters in the FitzHugh-Nagumo model (for plot details see Fig. 5 caption)

errors prevent estimation of parameters with lower than 3% error, as the solution is nearly discontinuous in this case (jumps between datapoints are  $\mathcal{O}(1)$ ).

### 3.2.4 Hindmarsh-Rose

The Hindmarsh-Rose model is used to emulate neuronal bursting and features 10 fundamental parameters which span 4 orders of magnitude (Hindmarsh and Rose 1984). Bursting behavior is observed in the first two solution components, while the third component represents slow neuronal adaptation with dynamics that are two orders of magnitude smaller in amplitude. Bursting produces steep gradients which render the dynamics numerically discontinuous at  $M = 128$  timepoints, while at  $M = 256$  there is at most one data point between peaks and troughs of bursts (see Fig. 8, upper left). Furthermore, cubic and quadratic nonlinearities lead to inaccuracies at high levels of noise. Thus, in a multitude of ways (multiple coefficient scales, multiple solution scales, steep gradients, higher-order nonlinearities, etc.) this is a challenging problem, yet an important one as it exhibits a canonical biological phenomenon. Figure 8 (lower left) shows that WENDy is robust to 2% noise when  $M \geq 256$ , robust to 5% noise when  $M \geq 512$ , and robust to 10% noise when  $M \geq 1024$ . It should be noted that since our noise model applies additive noise of equal variance to each component, relatively small noise renders the slowly-varying third component  $u_3$  unidentifiable (in fact, the noise ratio of only  $\mathbf{U}^{(3)}$  exceeds 100% when the total noise ratio is 10%).

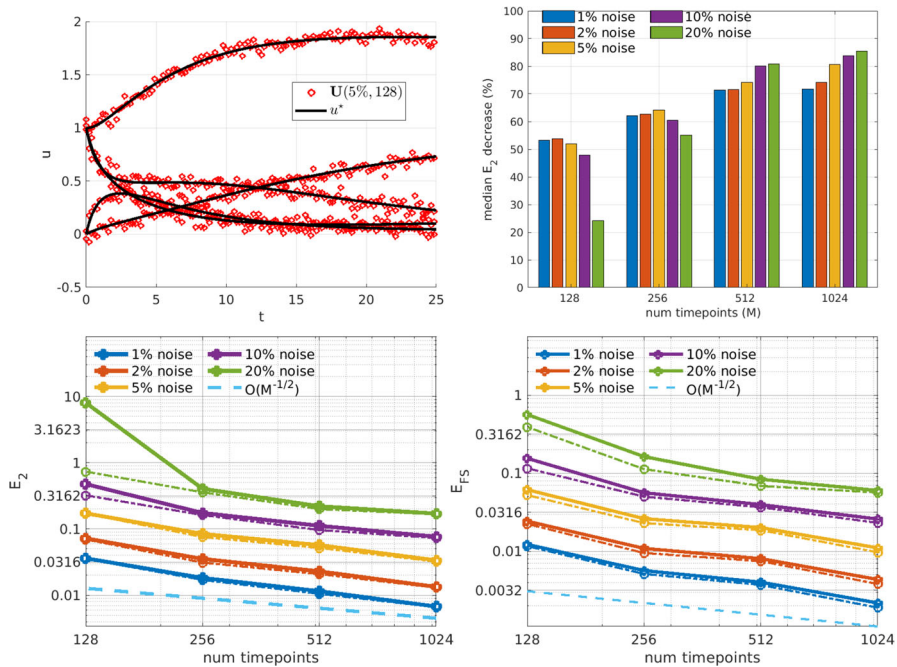


**Fig. 8** Hindmarsh-rose: Estimation of parameters in the Hindmarsh-Rose model (for plot details see Fig. 5 caption)

In the operable range of 1%–2% noise and  $M \geq 256$ , WENDy results in 70%–90% reductions in errors from the naive OLS solution, indicating that inclusion of the approximate covariance is highly beneficial under conditions which can be assumed to be experimentally relevant. We note that the forward simulation error here is not indicative of performance, as it will inevitably be large in all cases due to slight misalignment with bursts in the true data.

### 3.2.5 Protein Transduction Benchmark (PTB)

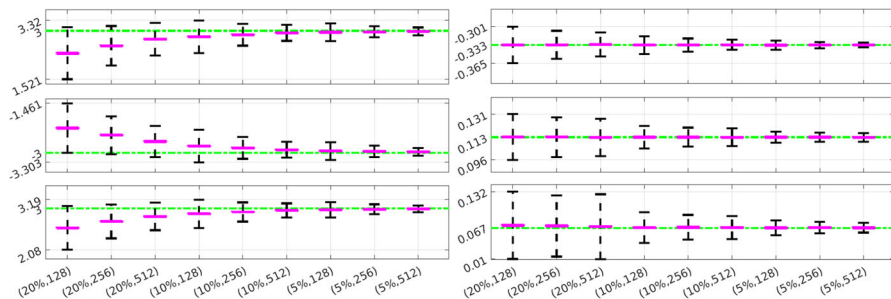
The PTB model is a five-compartment protein transduction model identified in Schoeberl et al. (2002) as a mechanism in the signaling cascade of epidermal growth factor (EGF). It was used in Vysheirsky and Girolami (2008) to compare between four other models, and has since served as a benchmark for parameter estimation studies in biochemistry (Macdonald and Husmeier 2015; Niu et al. 2016; Kirk et al. 2013). The nonlinearities are quadratic and sigmoidal, the latter category producing nontrivial transformations of the additive noise. WENDy estimates the 11 parameters with reasonable accuracy when 256 or more timepoints are available (Fig. 9), which is sufficient to result in forward simulation errors often much less than 10%. The benefit of using WENDy over the OLS solution is most apparent for  $M \geq 512$ , where the coefficient errors are reduced by at least 70%, leading to forward simulation errors less than 10%, even at 20% noise.



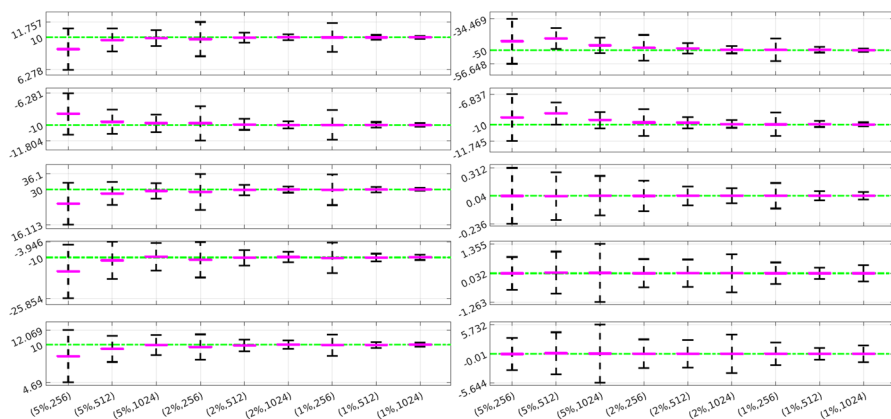
**Fig. 9** Protein transduction benchmark (PTB): Estimation of parameters in the PTB model (for plot details see Fig. 5 caption)

### 3.3 Parameter Uncertainties Using Learned Covariance

We now demonstrate how the WENDy methodology can be used to communicate uncertainty of the parameter estimates, and comment on the performance of the WENDy confidence intervals over repeated simulations. Figures 10 and 11 contain visualizations of average confidence intervals around each parameter in the FitzHugh-



**Fig. 10** FitzHugh-nagumo: Performance of WENDy for all estimated parameters. The true parameters are plotted in green, the purple lines indicate the average learned parameters over all experiments and the black lines represent the average 95% confidence intervals obtained by applying Eq. (16) using the average learned parameter covariance matrix  $\mathbf{S}$ . The  $x$ -axis indicates noise level and number of timepoints for each interval (Color figure online)



**Fig. 11** Hindmarsh-Rose: Performance of WENDy for all estimated parameters. See Fig. 10 for a description

Nagumo and Hindmarsh-Rose models, calculated over repeated simulations with specific noise levels and numbers of timepoints. For each combination of noise level and number of timepoints, we computed an average 95% confidence interval around the average learned parameter using Eq. (16) and the averaged covariance matrix<sup>14</sup>. As expected, increasing the number of timepoints and decreasing the noise level leads (on average) to more certainty in the learned parameters, while lower quality data leads on average to higher uncertainty. The ability to reliably assess uncertainty is useful not only for our understanding of the precision with which the method can estimate parameters, but also for designing most efficient experimental protocols (Keck and Bortz 2016), and assessing resulting uncertainty in the state predictions and decision functionals based on the fitted model (Elder et al. 2006).

One could also examine the off-diagonal correlations in  $\mathbf{S}$ . In Table 2 we show the average  $\mathbf{S}$  matrix for FitzHugh-Nagumo at 20% noise using 128 timepoints. This example illustrates a situation where parameter estimates for  $w_1$ ,  $w_2$ , and  $w_3$  tend to be highly correlated pairwise, and that an average dataset with this experimental setting does not provide much information to estimate separately each of these individual parameters very precisely. This may seem intuitive because these parameters' absolute values are equal, but the terms they correspond to are very different, so this insight would not necessarily be known *a priori*. Similarly, when looking at the confidence intervals in Fig. 10 (left column), we observe that  $w_1$ ,  $w_2$ ,  $w_3$  also exhibit the highest variance, meaning again that the uncertainty in these individual parameter estimates is high due to the data's reduced ability to support separate precise estimation of these parameters. This indicates that it may be possible to reduce the total variance of all states or decision functionals based on these parameters by incorporating their joint correlation structure. We can also observe that  $w_3$  exhibits almost no correlation with  $w_6$  on average, despite corresponding to the same term (albeit in different equations), and that coefficients in the second equation ( $w_4$ ,  $w_5$ ,  $w_6$ ) do not exhibit the same high

<sup>14</sup> Scripts are available at <https://github.com/MathBioCU/WENDy> to generate similar plots for the other examples.

**Table 2** Entries of average learned parameter covariance matrix  $\mathbf{S}$  matrix for Fitzhugh-Nagumo data with 20% noise and 128 timepoints, scaled to have 1's along the diagonal

	$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	$w_6$
$w_1$	1.000	-0.984	0.850	0.233	-0.202	-0.186
$w_2$		1.000	-0.813	-0.178	0.203	0.135
$w_3$			1.000	0.411	-0.306	0.003
$w_4$				1.000	-0.551	-0.136
$w_5$					1.000	-0.183
$w_6$						1.000

level of correlation as those in the first equation. We aim to explore these directions in a future work.

### 3.4 Comparison to Nonlinear Least Squares

We now briefly compare WENDy and forward solver-based nonlinear least squares (FSNLS) using walltime and relative coefficient error  $E_2$  as criteria. For nonlinear least-squares one must specify the initial conditions for the ODE solve (IC), a simulation method (SM), and an initial guess for the parameters ( $\mathbf{w}^{(0)}$ ). Additionally, stopping tolerances for the optimization method must be specified (Levenberg-Marquardt is used throughout). Optimal choices for each of these hyperparameters is an ongoing area of research. We have optimized FSNLS in ways that are unrealistic in practice in order to demonstrate the advantages of WENDy even when FSNLS is performing somewhat optimally in both walltime and accuracy. Our hyperparameter selections are collected in Table 3 and discussed below.

To remove some sources of error from FSNLS, we use the true initial conditions  $u(0)$  throughout, noting that these would not be available in practice. For the simulation method, we use state-of-the-art ODE solvers for each problem, namely for the stiff differential equations Fitzhugh-Nagumo and Hindmarsh-Rose we use MATLAB's `ode15s`, while for Lotka-Volterra and PTB we use `ode45`. In this way FSNLS is optimized for speed in each problem. We fix the relative and absolute tolerances of the solvers at  $10^{-6}$  in order to prevent numerical errors from affecting results without asking for excessive computations. In practice, the ODE tolerance, as well as the solver, must be optimized to depend on the noise in the data, and the relation between simulation errors and parameters errors in FSNLS is an on-going area of research (Nardini and Bortz 2019).

Due to the non-convexity of the loss function in FSNLS, choosing a good initial guess  $\mathbf{w}^{(0)}$  for the parameters  $\mathbf{w}^*$  is crucial. For comparison, we use two strategies. The first strategy (simply labeled FSNLS in Figs. 12, 13, 14 and 15), consists of running FSNLS on five initial guesses, where each parameter is sampled i.i.d from a uniform distribution, i.e., for the  $i$ th parameter,

$$\mathbf{w}_i^{(0)} \sim \mathbf{w}_i^* + U([- \sigma/2, \sigma/2])$$

**Table 3** Hyperparameters for the FSNLS algorithm

IC	Simulation method	$\mathbf{w}^{(0), \text{batch}}$	$\mathbf{w}^{(0), \text{WENDy}}$	Max. evals	Max. iter	Min. step
$u^*(0)$	L-V, PTB: ode45 FH-N, H-R: ode15s (abs/rel tol= $10^{-6}$ )	$\mathbf{w}^{(0)} \sim U(\mathbf{w}^*, \sigma)$ , best out of 5	$\mathbf{w}^{(0)} = \hat{\mathbf{w}}$	2000	500	$10^{-8}$

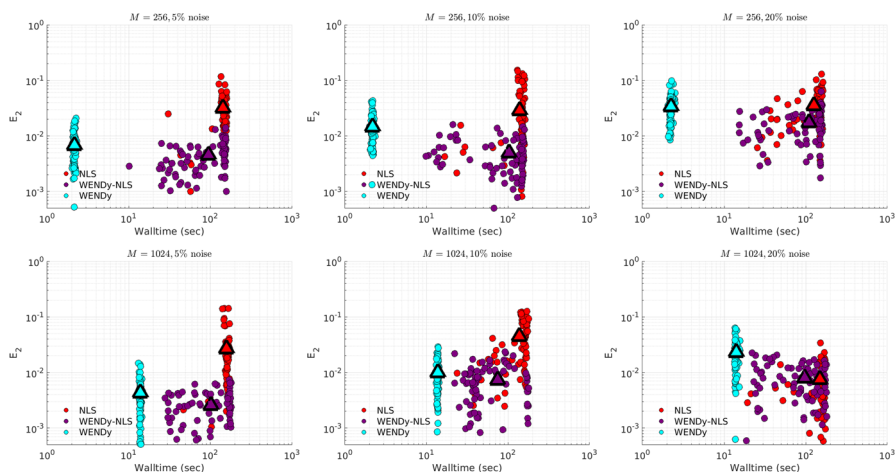
and keeping only the best-performing result. Since the sign of coefficients greatly impacts the stability of the ODE, we take the standard deviations to be

$$\sigma_j = 0.25|\mathbf{w}_j^*| \quad (25)$$

so that initial guesses always have the correct sign but with approximately 25% error from the true coefficients. (For cases like Hindmarsh-Rose, this implies that the small coefficients in  $\mathbf{w}^*$  are measured to high accuracy relative to the large coefficients.) In practice, one would not have the luxury of selecting the lowest-error result of five independent trials of FSNLS, however it may be possible to combine several results to boost performance.

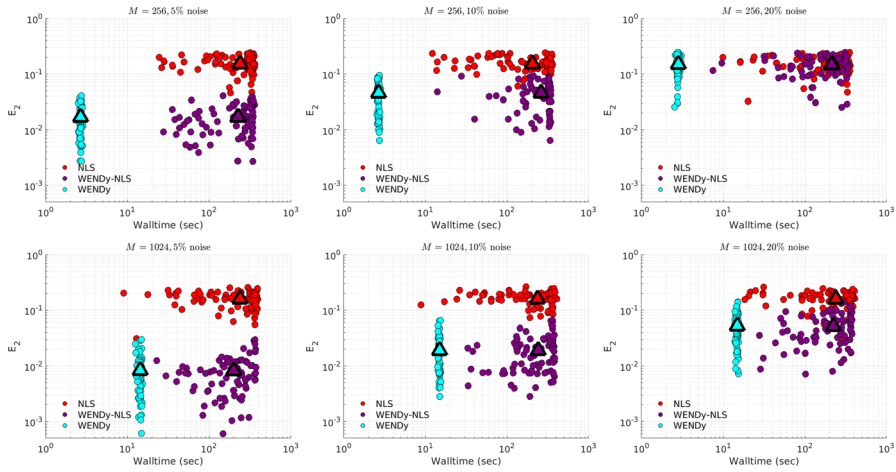
For the second initial guess strategy we set  $\mathbf{w}^{(0)} = \hat{\mathbf{w}}$ , the output from WENDy (labeled WENDy-FSNLS in Figs. 12, 13, 14 and 15). In almost all cases, this results in an increase in accuracy, and in many cases, also a decrease in walltime.

Figures 12, 13, 14 and 15 display comparisons between FSNLS, WENDy-FSNLS, and WENDy for Lotka-Volterra, FitzHugh-Nagumo, Hindmarsh-Rose, and PTB models. In general, we observe that WENDy provides significant decreases in walltime and

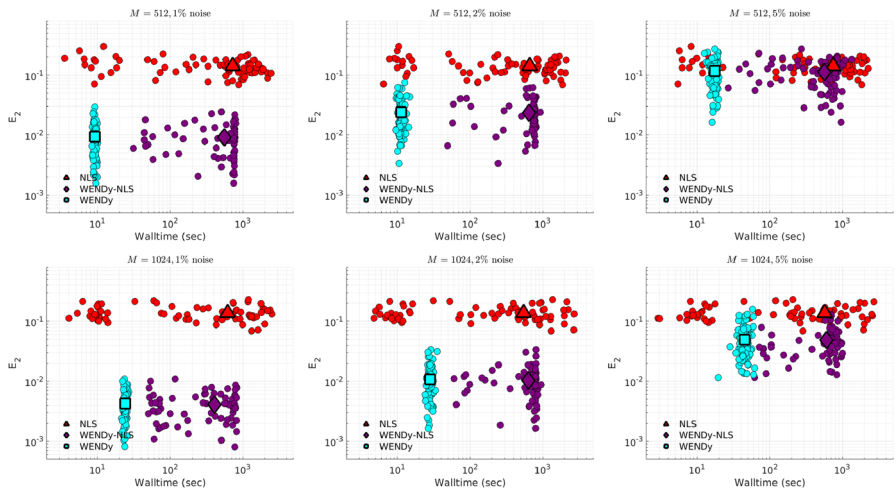


**Fig. 12** Comparison between FSNLS, WENDy-FSNLS, and WENDy for the Lotka-Volterra model. Left to right: noise levels {5%, 10%, 20%}. Top: 256 timepoints, bottom: 1024 timepoints. We note that the  $M = 1024$  with 20% noise figure on the lower right suggests that WENDy results in slightly higher errors than the FSNLS. This is inconsistent with all other results in this work and appears to be an outlier. Understanding the source of this discrepancy is a topic or future work



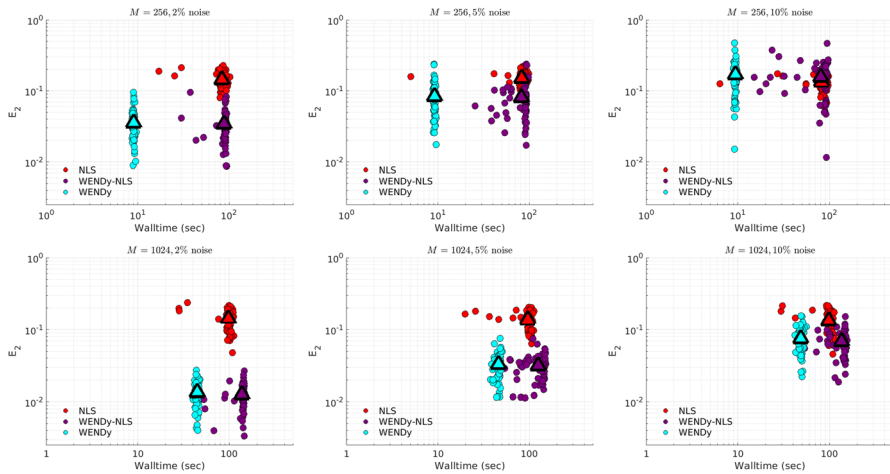


**Fig. 13** Comparison between FSNLS, Wendy-FSNLS, and WENDy for the FitzHugh-Nagumo model. Left to right: noise levels {5%, 10%, 20%}. Top: 256 timepoints, bottom: 1024 timepoints



**Fig. 14** Comparison between FSNLS, Wendy-FSNLS, and WENDy for the Hindmarsh-Rose model. Left to right: noise levels {1%, 2%, 5%}. Top: 512 timepoints, bottom: 1024 timepoints

modest to considerable increases in accuracy compared to the FSNLS solution. Due to the additive noise structure of the data, this is surprising because FSNLS corresponds to (for normally distributed measurement errors) a maximum likelihood estimation, while WENDy only provides a first order approximation to the statistical model. At lower resolution and higher noise (top right plot in Figs. 12, 13, 14 and 15), all three methods are comparable in accuracy, and WENDy decreases the walltime by two orders of magnitude. In several cases, such as Lotka-Volterra Fig. 12, the WENDy-FSNLS solution achieves a lower error than both WENDy and FSNLS, and improves on the speed of FSNLS. For Hindmarsh-Rose, even with high-resolution data and low noise (bottom left plot of Fig. 14), FSNLS is unable to provide an accurate solution



**Fig. 15** Comparison between FSNLS, WENDy-FSNLS, and WENDy for the PTB model. Left to right: noise levels {2%, 5%, 10%}. Top: 256 timepoints, bottom: 1024 timepoints

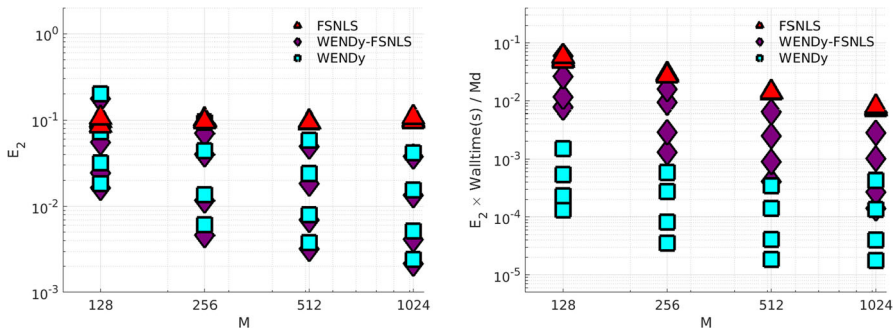
( $E_2 \approx 0.2$ ), while WENDy and WENDy-FSNLS result in  $E_2 \approx 0.005$ . The clusters of FSNLS runs in Fig. 14 with walltimes  $\approx 10$  seconds correspond to local minima, a particular weakness of FSNLS, while the remaining runs have walltimes on the order of 20 min, compared to 10–30 s WENDy. We see a similar trend in  $E_2$  for the PTB model (Fig. 15), with  $E_2$  rarely dropping below 10%, however in this case FSNLS runs in a more reasonable amount of time, taking only  $\approx 100$  s. The WENDy solution offers speed and error reductions. For high-resolution data ( $M = 1024$ ), WENDy runs in 40–50 s on PTB data due to the impact of  $M$  and  $d$ , the number of ODE compartments (here  $d = 5$ ), on the computational complexity. It is possible to reduce this using more a sophisticated implementation (in particular, symbolic computations are used to take gradients of generic functions, which could be precomputed).

Finally, the aggregate performance of WENDy, WENDy-FSNLS, and FSNLS is reported in Fig. 16, which reiterates the trends identified in the previous Figures. Firstly, WENDy provides significant accuracy and walltime improvements over FSNLS. It is possible that FSNLS results in lower error for very small sample sizes (see  $M = 128$  results in the left plot), although this comes at a much higher computational cost. Secondly, WENDy-FSNLS provides similar accuracy improvements over FSNLS and improves the walltime per datapoint score, suggesting that using WENDy as an initial guess may alleviate the computational burden in cases where FSNLS is competitive.

## 4 Concluding Remarks

In this work, we have proposed the Weak-form Estimation of Nonlinear Dynamics (WENDy) method for directly estimating model parameters, without relying on forward solvers. The essential feature of the method involves converting the strong form representation of a model to its weak form and then substituting in the data and solving





**Fig. 16** Average performance of FSNLS, WENDy-FSNLS, and WENDy over Lotka-Volterra, FitzHugh-Nagumo, Hindmarsh-Rose and PTB for noise ratios  $\sigma_{NR} \in \{0.01, 0.02, 0.05, 0.1\}$ . To account for scaling between examples, the geometric mean across the four examples is reported in each plot. Left: average relative coefficient error  $E_2$  vs. number of timepoints  $M$ ; right: relative coefficient error  $E_2$  multiplied by walltime per datapoint vs.  $M$ . In each case, increasing noise levels  $\sigma_{NR}$  correspond to increasing values along the y-axis. Both plots suggest that WENDy and WENDy-FSNLS each provide accuracy and walltime improvements over FSNLS with best-of-five random initial parameter guesses

a regression problem for the parameters. The method is robust to substantial amounts of noise, and in particular to levels frequently seen in biological experiments.

As mentioned above, the idea of substituting data into the weak form of an equation followed by a least squares solve for the parameters has existed since at least the mid 1950's (Shinbrot 1954). However, due to their performance, FSNLS-based methods have dominated and are ubiquitous in the parameter estimation literature and available software. The disadvantage of FSNLS is that fitting using repeated forward solves comes at a substantial computational cost and with unclear dependence on the initial guess and hyperparameters (in both the solver and the optimizer). Several researchers over the years have created direct parameter estimation methods (that do not rely on forward solves), but they have historically included some sort of data smoothing step. The primary issue with this is that projecting the data onto a spline basis (for example) represents the data using a basis which does not solve the original equation<sup>15</sup>. Importantly, that error propagates to the error in the parameter estimates. However, we note that the WENDy framework introduced here is able to encapsulate previous works that incorporate smoothing, namely by including the smoothing operator in the covariance matrix  $\hat{\mathbf{C}}$ .

The conversion to the weak form is essentially a weighted integral transform of the equation. As there is no projection onto a non-solution based function basis, the weak-form approach bypasses the need to estimate the true solution to directly estimate the parameters.

The main message of this work is that weak-form-based direct parameter estimation offers intriguing advantages over FSNLS-based methods. In almost all the examples shown in this work and in particular for larger dimensional systems with high noise, the WENDy method is faster and more accurate by orders of magnitude. In rare cases where an FSNLS-based approach yields higher accuracy, WENDy can be used as an efficient method to identify a good initial guess for parameters.

<sup>15</sup> This is a problem WENDy does not suffer from as there is no pre-smoothing of the data.

**Acknowledgements** The authors would like to thank Dr. Michael Zager (Pfizer) and Dr. Clay Thompson (SAS) for offering insight into the state of the art parameter estimation methods used in industry.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Banks HT, Kunisch K (1989) Estimation techniques for distributed parameter systems, systems and control: foundations and applications, vol 1. Birkhäuser Boston, Boston, MA
- Bellman R (1969) A new method for the identification of systems. *Math Biosci* 5(1–2):201–204. [https://doi.org/10.1016/0025-5564\(69\)90042-X](https://doi.org/10.1016/0025-5564(69)90042-X)
- Bertsimas D, Gurnee W (2023) Learning sparse nonlinear dynamics via mixed-integer optimization. *Nonlinear Dyn* 111:6585–6604. <https://doi.org/10.1007/s11071-022-08178-9>
- Bollerslev T, Wooldridge JM (1992) Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econom Rev* 11(2):143–172. <https://doi.org/10.1080/07474939208800229>
- Bortz DM (2006) Accurate model selection computations. *J Crit Care* 21(4):359
- Brunel NJB (2008) Parameter estimation of ODE's via nonparametric estimators. *Electron J Stat* 2:1242–1267. <https://doi.org/10.1214/07-EJS132>
- Brunel NJB, Clairon Q, d'Alché-Buc F (2014) Parametric estimation of ordinary differential equations with orthogonality conditions. *J Am Stat Assoc* 109(505):173–185. <https://doi.org/10.1080/01621459.2013.841583>
- Brunton SL, Proctor JL, Kutz JN (2016) Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc Natl Acad Sci* 113(15):3932–3937. <https://doi.org/10.1073/pnas.1517384113>
- Calderhead B, Girolami M, Lawrence ND (2008) Accelerating Bayesian inference over nonlinear differential equations with gaussian processes. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) *Advances in neural information processing systems*, vol 21. Curran Associates Inc
- Dattner I (2021) Differential equations in data analysis. *WIREs Comp Stat* 13(6):e1534. <https://doi.org/10.1002/wics.1534>
- Dattner I, Miller E, Petrenko M, Kadouri DE, Jurkevitch E, Huppert A (2017) Modelling and parameter inference of predator-prey dynamics in heterogeneous environments using the direct integral approach. *J R Soc Interface* 14(126):20160525. <https://doi.org/10.1098/rsif.2016.0525>
- Ding AA, Wu H (2014) Estimation of ordinary differential equation parameters using constrained local polynomial regression. *Stat Sin* 24(4):1613–1631. <https://doi.org/10.5705/ss.2012.304>
- Duistermaat J, Kolk J (2010) *Distributions kernels*. Birkhäuser Boston, Boston
- Elder BD, Dukic VM, Dwyer G (2006) Uncertainty in predictions of disease spread and public health responses to bioterrorism and emerging diseases. *Proc Natl Acad Sci* 103(42):15693–15697. <https://doi.org/10.1073/pnas.0600816103>
- Fasel U, Kutz JN, Brunton BW, Brunton SL (2021) Ensemble-SINDy: robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proc R Soc A* 478(2260):20210904. <https://doi.org/10.1098/rspa.2021.0904>
- FitzHugh R (1961) Impulses and physiological states in theoretical models of nerve membrane. *Biophys J* 1(6):445–466. [https://doi.org/10.1016/S0006-3495\(61\)86902-6](https://doi.org/10.1016/S0006-3495(61)86902-6)
- Fornberg B (1988) Generation of finite difference formulas on arbitrarily spaced grids. *Math Comput* 51(184):699–699. <https://doi.org/10.1090/S0025-5718-1988-0935077-0>
- Greenberg H (1951) A survey of methods for determining stability parameters of an airplane from dynamics flight measurements. Technical Report NACA TN 2340, Ames Aeronautical Laboratory, Moffett Field, CA

- Gurevich DR, Reinbold PAK, Grigoriev RO (2019) Robust and optimal sparse regression for nonlinear PDE models. *Chaos* 29(10):103113. <https://doi.org/10.1063/1.5120861>
- Hindmarsh JL, Rose RM (1984) A model of neuronal bursting using three coupled first order differential equations. *Proc R Soc Lond B* 221(1222):87–102. <https://doi.org/10.1098/rspb.1984.0024>
- Jorgensen M (2012) Iteratively reweighted least squares. In: El-Shaarawi AH, Piegorsch WW (eds) *Encyclopedia of environmetrics*, 1st edn. Wiley
- Kaptanoglu A, de Silva B, Fasel U, Kaheman K, Goldschmidt A, Callahan J, Delahun C, Nicolaou Z, Champion K, Loiseau JC, Kutz J, Brunton S (2022) PySINDy: a comprehensive Python package for robust sparse system identification. *JOSS* 7(69):3994. <https://doi.org/10.21105/joss.03994>
- Keck DD, Bortz DM (2016) Generalized sensitivity functions for size-structured population models. *J Inverse Ill-Posed Probl* 24(3):309–321. <https://doi.org/10.1515/jiip-2014-0041>
- Khanmohamadi O, Xu D (2009) Spatiotemporal system identification on nonperiodic domains using Chebyshev spectral operators and system reduction algorithms. *Chaos* 19(3):033117. <https://doi.org/10.1063/1.3180843>
- Kirk P, Thorne T, Stumpf MP (2013) Model selection in systems and synthetic biology. *Curr Opin Biotechnol* 24(4):767–774. <https://doi.org/10.1016/j.copbio.2013.03.012>
- Lax PD, Milgram AN (1955) IX parabolic equations, *annals of mathematical studies*, vol 33. Princeton University Press, pp 167–190
- Liang H, Wu H (2008) Parameter estimation for differential equation models using a framework of measurement error in regression models. *J Am Stat Assoc* 103(484):1570–1583. <https://doi.org/10.1198/016214508000000797>
- Ljung L (1999) *System identification: theory for the user*, 2nd edn. Prentice hall information and system sciences series. Prentice Hall PTR, Upper Saddle River, NJ
- Ljung L (2017) *System identification*. In: Webster J (ed) *Wiley encyclopedia of electrical and electronics engineering*. Wiley, Hoboken, NJ, USA, pp 1–19
- Loeb J, Cahen GM (1965a) Extraction a partir des enregistrements de mesures, des parametres dynamiques d'un systeme. *Automatisme* 8:479–486
- Loeb J, Cahen GM (1965b) More about process identification. *IEEE Trans Autom Control* 10(3):359–361. <https://doi.org/10.1109/TAC.1965.1098172>
- Lotka AJ (1978) The growth of mixed populations: two species competing for a common food supply. In: Scudo F, Ziegler J (eds) *The golden age of theoretical ecology: 1923–1940*, vol 22. Springer, Berlin, Heidelberg, pp 274–286
- Macdonald B, Husmeier D (2015) Computational inference in systems biology. In: Ortu F, Rojas I (eds) *Bioinformatics and biomedical engineering*, vol 9044. Springer International Publishing, Cham, pp 276–288
- Martina-Perez S, Simpson MJ, Baker RE (2021) Bayesian uncertainty quantification for data-driven equation learning. *Proc R Soc A* 477(2254):20210426. <https://doi.org/10.1098/rspa.2021.0426>
- Messenger DA, Bortz DM (2021a) Weak SINDy for partial differential equations. *J Comput Phys* 443:110525. <https://doi.org/10.1016/j.jcp.2021.110525>
- Messenger DA, Bortz DM (2021b) Weak SINDy: galerkin-based data-driven model selection. *Multiscale Model Simul* 19(3):1474–1497. <https://doi.org/10.1137/20M1343166>
- Messenger DA, Bortz DM (2022a): Asymptotic consistency of the WSINDy algorithm in the limit of continuum data. [arXiv:2211.16000](https://arxiv.org/abs/2211.16000) (submitted)
- Messenger DA, Bortz DM (2022b) Learning mean-field equations from particle data using WSINDy. *Phys D* 439:133406. <https://doi.org/10.1016/j.physd.2022.133406>
- Messenger DA, Dall'Anese E, Bortz DM (2022a) Online Weak-form Sparse Identification of Partial Differential Equations. In: *Proceedings of Third Mathematics Science Machine Learning Conference, Proceedings of Machine Learning Research*, vol. 190, pp. 241–256. PMLR
- Messenger DA, Wheeler GE, Liu X (2022b) Learning anisotropic interaction rules from individual trajectories in a heterogeneous cellular population. *J R Soc Interface* 19(195):20220412. <https://doi.org/10.1098/rsif.2022.0412>
- Nardini JT, Bortz DM (2019) The influence of numerical error on parameter estimation and uncertainty quantification for advective PDE models. *Inverse Probl* 35(6):065003. <https://doi.org/10.1088/1361-6420/ab10bb>
- Nicolaou ZG, Huo G, Chen Y, Brunton SL, Kutz JN (2023) Data-driven discovery and extrapolation of parameterized pattern-forming dynamics. [arXiv:2301.02673](https://arxiv.org/abs/2301.02673)

- Niu M, Rogers S, Filippone M, Husmeier D (2016) Fast Inference in Nonlinear Dynamical Systems using Gradient Matching. In: Proceedings of the 33rd International Conference on Machine Learning, vol. 48, pp. 1699–1707. PMLR
- Pantazis Y, Tsamardinos I (2019) A unified approach for sparse dynamical system inference from temporal measurements. *Bioinformatics* 35(18):3387–3396. <https://doi.org/10.1093/bioinformatics/btz065>
- Perdreauville FJ, Goodson RE (1966) Identification of systems described by partial differential equations. *J Basic Eng* 88(2):463–468. <https://doi.org/10.1115/1.3645880>
- Poyton A, Varzi M, McAuley K, McLellan P, Ramsay J (2006) Parameter estimation in continuous-time dynamic models using principal differential analysis. *Comput Chem Eng* 30(4):698–708. <https://doi.org/10.1016/j.compchemeng.2005.11.008>
- Preisig H (1993) Theory and application of the modulating function method—I. Review and theory of the method and theory of the spline-type modulating functions. *Comput Chem Eng* 17(1):1–16. [https://doi.org/10.1016/0098-1354\(93\)80001-4](https://doi.org/10.1016/0098-1354(93)80001-4)
- Ramsay JO, Hooker G, Campbell D, Cao J (2007) Parameter estimation for differential equations: a generalized smoothing approach. *J R Stat Soc Ser B Stat Methodol* 69(5):741–796. <https://doi.org/10.1111/j.1467-9868.2007.00610.x>
- Reinbold PAK, Gurevich DR, Grigoriev RO (2020) Using noisy or incomplete data to discover models of spatiotemporal dynamics. *Phys Rev E* 101(1):010203. <https://doi.org/10.1103/PhysRevE.101.010203>
- Rudy SH, Brunton SL, Proctor JL, Kutz JN (2017) Data-driven discovery of partial differential equations. *Sci Adv* 3(4):e1602614. <https://doi.org/10.1126/sciadv.1602614>
- Sangalli LM (2021) Spatial regression With partial differential equation regularisation. *Int Stat Rev* 89(3):505–531. <https://doi.org/10.1111/insr.12444>
- Schaeffer H, McCalla SG (2017) Sparse model selection via integral terms. *Phys Rev E* 96(2):023302. <https://doi.org/10.1103/PhysRevE.96.023302>
- Schoeberl B, Eichler-Jonsson C, Gilles ED, Müller G (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat Biotechnol* 20(4):370–375. <https://doi.org/10.1038/nbt0402-370>
- Schwartz L (1950) *Théorie Des distributions*, vol I. Hermann et Cie, Paris, France
- Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). *Biometrika* 52(3–4):591–611. <https://doi.org/10.1093/biomet/52.3-4.591>
- Shinbrot M (1954) On the analysis of linear and nonlinear dynamical systems for transient-response data. Technical Report NACA TN 3288, Ames Aeronautical Laboratory, Moffett Field, CA
- Van Huffel S, Lemmerling P (eds) (2002) *Total least squares and errors-in-variables modeling: analysis algorithms and applications*. Springer, Netherlands, Dordrecht
- Varah JM (1982) A spline least squares method for numerical parameter estimation in differential equations. *SIAM J Sci Stat Comput* 3(1):28–46. <https://doi.org/10.1137/0903003>
- Vysheirsky V, Girolami MA (2008) Bayesian ranking of biochemical system models. *Bioinformatics* 24(6):833–839. <https://doi.org/10.1093/bioinformatics/btm607>
- Wang H, Zhou X (2021) Explicit estimation of derivatives from data and differential equations by Gaussian process regression. *Int J Uncertain Quantif* 11(4):41–57. <https://doi.org/10.1615/Int.J.UncertaintyQuantification.2021034382>
- Wang Z, Huan X, Garikipati K (2019) Variational system identification of the partial differential equations governing the physics of pattern-formation: Inference under varying fidelity and noise. *Comput Methods Appl Mech Eng* 356:44–74. <https://doi.org/10.1016/j.cma.2019.07.007>
- Wenk P, Abbati G, Osborne MA, Schölkopf B, Krause A, Bauer S (2020) ODIN: ODE-informed regression for parameter and state inference in time-continuous dynamical systems. *AAAI* 34(04):6364–6371. <https://doi.org/10.1609/aaai.v34i04.6106>
- Xu D, Khanmohamadi O (2008) Spatiotemporal system reconstruction using Fourier spectral operators and structure selection techniques. *Chaos* 18(4):043122. <https://doi.org/10.1063/1.3030611>
- Yang S, Wong SWK, Kou SC (2021) Inference of dynamic systems from noisy and sparse data via manifold-constrained Gaussian processes. *Proc Natl Acad Sci USA* 118(15):e2020397118. <https://doi.org/10.1073/pnas.2020397118>
- Zhang N, Nanshan M, Cao J (2022) A Joint estimation approach to sparse additive ordinary differential equations. *Stat Comput* 32(5):69. <https://doi.org/10.1007/s11222-022-10117-y>