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

David M. Bortz · Daniel A. Messenger · Vanja Dukic

Abstract We introduce the Weak-form Estimation of Nonlinear Dynamics (WENDy) method for estimating model parameters for non-linear systems of ODEs. 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^∞ bump functions of varying support sizes.

We demonstrate that WENDy is a highly robust and efficient method for parameter inference in differential equations. 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.

We illustrate the method and its performance 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

Mathematics Subject Classification (2000) 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

David M. Bortz
Department of Applied Mathematics
University of Colorado, Boulder, CO 80309-0526
Tel.: +1 303-492-7569
Fax: +1 303-492-4066
E-mail: david.bortz@colorado.edu

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 Section 2.1, we introduce the idea of weak-form parameter estimation, including a simple algorithm to illustrate the idea. In Section 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 Section 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 Section 2.3, we present a strategy for computing an orthogonal set of test functions that facilitate a successful weak-form implementation. In Section 3 we illustrate the performance of WENDy using five common mathematical models from the biological sciences and in Section 4 we offer some concluding remarks.

1.1 Background

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

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

where the function $u : \mathbb{R} \to \mathbb{R}^d$ is a solution to a differential equation model[1]

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

The ODE system in equation 2 is parameterized by $\mathbf{w} \in \mathbb{R}^J$, the vector of J true parameters which are to be estimated by $\widehat{\mathbf{w}}$. 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 equation (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 [25, 26] and, for those interested in a more theoretical perspective, to the monograph by Banks and Kunisch [1].

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 [38] for an illustration with PDE models and [5] for an example with ODE and DDE models. There is no widespread convention on detection of 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.

---

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

Perhaps more importantly, the choice of the initial candidate parameter vector can have a huge impact upon the final estimate, given that nonlinear least squares cost functions frequently have multiple local minima in differential equations applications. There are several algorithms designed to deal with the multi-modality, such as particle swarm optimization [4] and simulated annealing [54]; however, all come at the cost of additional forward solves and unclear dependence on the hyperparameters used in the solver and optimization algorithms.

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 equation (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 equation (2) – i.e., via a gradient matching criteria. Indeed, Richard Bellman proposed exactly this strategy in 1969 [2]. There have been similar ideas in the literature of chemical and aerospace engineering, which can be traced back even further [16,42]. 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 [9,31,57,59,61] and collocations projecting onto a polynomial or spline basis [6,24,43,45,55,62]. For example, Yang et al. [61], 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. Ramsey et al. [45] 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 [24] 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 [12] by using local polynomial regression instead of the pseudo-least squares estimator used in [24].

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 [60] and Chebyshev operators [21]. 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 <u>weak form</u> of the equation. The analysis of the weak form of a model was pioneered by Lax and Milgram in the 1950's for relaxing smoothness requirements on unique solutions to parabolic systems in Hilbert spaces [23]. 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 [44] for a good historical overview). Briefly, in 1954, Shinbrot created a proto-weak-form parameter inference method, called the Equations Of Motion (EOM) method [52]. 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 ν and n. In 1965, Loeb and Cahen [27,28] 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 [33], a critical step in obtaining robust and accurate parameter estimation is to use <u>highly</u> smooth test 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^∞ bump functions in WENDy (see Section 2.3).

In the statistics literature, there are several examples of using integral or weak-form equations. Dattner et al. [11] illustrate an integral-based approach and Dattner's 2021 review [10] 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 [7, 48]. 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 [8] is built upon direct parameter estimation methods, there are also several relevant contributions from this literature. Schaeffer and McCalla [49] 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 [17, 32, 33, 41, 41, 58]. The weak form is now even implemented in the PySINDy code [20] which is actively developed by the authors of the original SINDy papers [8, 47]. However, we do note that the Weak SINDy in PySINDy is based on an early weak form implementation (proposed in [17, 46]). A more recent implementation with autotuned hyperparameters can be found at https://github.com/MathBioCU/WSINDy_ODE for ODEs [33] and https://github.com/MathBioCU/WSINDy_PDE for PDEs [32].

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 [33], PDEs [32], interacting particle systems of the first [35] and second [37] order, and online streaming [36]. 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 [32]. Lastly we have also studied the theoretical convergence properties for WSINDy in the continuum data limit [34]. Among the results are a description of a broad class of models for which the asymptotic limit of continuum data can overcome <u>any</u> noise level to yield both an accurately learned equation and a correct parameter estimate (see [34] 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 used in WSINDy. We thus denote the method presented in this paper as the Weak-form Estimation of Nonlinear Dynamics (WENDy) method.

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

2.1 Weak-form Esimation with Ordinary Least Squares

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

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

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) $\Theta(u) := [\, f_1(u) \ f_2(u) \ \cdots \ f_J(u) \,]$ where $f_j : \mathbb{R}^d \to \mathbb{R}$, and the matrix of unknown parameters $W \in \mathbb{R}^{J \times d}$. We consider a C^∞ test function ϕ compactly supported in the time interval $[0, T]$ (e.g. $\phi \in C_c^\infty([0, T])$), multiply both sides of (3) by ϕ, and integrate over 0 to T. Via integration by parts we obtain

$$\phi(T)u(T) - \phi(0)u(0) - \int_0^T \dot{\phi}u\, \mathsf{d}t = \int_0^T \phi\Theta(u)W\, \mathsf{d}t.$$

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

$$-\int_0^T \dot{\phi}u\, \mathsf{d}t = \int_0^T \phi\Theta(u)W\, \mathsf{d}t. \tag{4}$$

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

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

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

$$-\dot{\Phi}_k\mathbf{u} \approx \Phi_k\Theta(\mathbf{u})W, \tag{5}$$

where

$$\Phi_k := \left[\, \phi_k(t_0) \middle| \cdots \middle| \phi_k(t_M) \,\right] \mathbf{Q}, \qquad \dot{\Phi}_k := \left[\, \dot{\phi}_k(t_0) \middle| \cdots \middle| \dot{\phi}_k(t_M) \,\right] \mathbf{Q}$$

and ϕ_k is a test function centered at timepoint t_{m_k}. To account for proper scaling, in computations we normalize each test function ϕ_k to have unit ℓ_2-norm, or $\sum_{m=0}^M \phi_k^2(t_m) = 1$.

The $\mathbf{Q}$ matrix contains the quadrature weights on the diagonal. In this work we use the composite Trapezoidal rule[2] for which the matrix is

$$\mathbf{Q} := \mathsf{diag}(^{\Delta t}/_2, \Delta t, \ldots, \Delta t, {}^{\Delta t}/_2) \in \mathbb{R}^{(M+1) \times (M+1)}.$$

We defer full consideration of the integration error until Section 2.3 but note that in the case of a non-uniform timegrid, $\mathbf{Q}$ 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

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

---

[2] 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.

where "vec" vectorizes a matrix,

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

and where $\mathbf{U}$ represents the data. The integration matrices are

$$\Phi = \begin{bmatrix} \Phi_1 \\ \vdots \\ \Phi_K \end{bmatrix} \in \mathbb{R}^{K \times (M+1)} \quad \text{and} \quad \dot{\Phi} = \begin{bmatrix} \dot{\Phi}_1 \\ \vdots \\ \dot{\Phi}_K \end{bmatrix} \in \mathbb{R}^{K \times (M+1)}.$$

The ordinary least squares (OLS) solution to (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 Θ_i for each variable $i = 1 \ldots, d$ and use regression for state i to solve for a vector $\widehat{\mathbf{w}}_i$ of parameters (instead of a matrix of parameters 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 $\{\widehat{W}\}$
>
> // Solve Ordinary Least Squares Problem
> 1 $\mathbf{G} \leftarrow \Phi\Theta(\mathbf{U})$
> 2 $\mathbf{B} \leftarrow -\dot{\Phi}\mathbf{U}$
> 3 $\widehat{W} \leftarrow (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{B}$

The OLS solution has respectable performance in some cases, but in general there is a clear need for improvement upon OLS. In particular, we note that (6) is <u>not</u> 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.[3] While a full and rigorous analysis of the statistical properties of weak-form estimation is beyond the scope of this article[4], here we will present several formal derivations aimed at improving the accuracy of parameter estimation. 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 acknowledge that the regression problem does not fit within the framework of ordinary least squares (see Figure 2) 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 $\mathbf{w}^\star$ and let $u^\star := u(t; \mathbf{w}^\star)$ and $\Theta^\star := \Theta(u^\star)$. We also assume that measurements of the system are noisy, so that at each timepoint t all states are observed with additive noise

$$U(t) = u^\star(t) + \varepsilon(t) \tag{7}$$

---

[3] The EiV problem with standard additive i.i.d. Gaussian measurement errors is known as a Total Least Squares (TLS) problem in applied and computational mathematics. The literature of EiV and TLS is very similar, but TLS problems are a subset of EiV problems. We direct the interested reader to [53] for more information.

[4] See our work in [34] for an investigation of the asymptotic consistency in the limit of continuum data.

where each element of $\varepsilon(t)$ is i.i.d. $\mathcal{N}(0, \sigma^2)$.[5] 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 $\mathbf{U} := \mathbf{u}^\star + \boldsymbol{\varepsilon} \in \mathbb{R}^{(M+1) \times d}$ and vector of parameters to be identified $\mathbf{w} \in \mathbb{R}^{Jd}$. The use bolded variables to represent evaluation 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

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

where we redefine

$$\mathbf{G} := [\mathbb{I}_d \otimes (\Phi \Theta(\mathbf{U}))],$$
$$\mathbf{b} := -\mathsf{vec}(\dot{\Phi}\mathbf{U}).$$

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

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

where

$$\mathbf{G}^\star := [\mathbb{I}_d \otimes (\Phi \Theta(\mathbf{u}^\star))],$$
$$\mathbf{b} := \underbrace{-\mathsf{vec}(\dot{\Phi}\mathbf{u}^\star)}_{\mathbf{b}^\star} + \underbrace{-\mathsf{vec}(\dot{\Phi}\boldsymbol{\varepsilon})}_{\mathbf{b}^{\boldsymbol{\varepsilon}}}.$$

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 Section 2.3).

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

$$\begin{aligned}
\mathbf{e}_\Theta - \mathbf{b}^{\boldsymbol{\varepsilon}} &= (\mathbf{G} - \mathbf{G}^\star)\mathbf{w} + \mathsf{vec}(\dot{\Phi}\boldsymbol{\varepsilon}) \\
&= \left[\mathbb{I}_d \otimes \left(\Phi\left(\Theta(\mathbf{U}) - \Theta(\mathbf{U} - \boldsymbol{\varepsilon})\right)\right)\right]\mathbf{w} + \left[\mathbb{I}_d \otimes \dot{\Phi}\right]\mathsf{vec}(\boldsymbol{\varepsilon}) \\
&= \mathbf{L}_{\mathbf{w}}\mathsf{vec}(\boldsymbol{\varepsilon}) + \mathbf{h}(\mathbf{U}, \mathbf{w}, \boldsymbol{\varepsilon})
\end{aligned} \tag{9}$$

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 (9) is

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

---

[5] 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.

where "mat" is the matricization operation and $\mathbf{P}$ is a permutation matrix such that $\mathbf{P}\mathsf{vec}(\boldsymbol{\varepsilon}) = \mathsf{vec}(\boldsymbol{\varepsilon}^T)$. The matrix $\nabla\Theta$ contains derivatives of the features

$$
\nabla\Theta := \left[
\begin{array}{c|c|c}
\nabla f_1(\mathbf{U}_0) & & \\
& \ddots & \\
& & \nabla f_1(\mathbf{U}_M) \\
\hline
& \vdots & \\
\hline
\nabla f_J(\mathbf{U}_0) & & \\
& \ddots & \\
& & \nabla f_J(\mathbf{U}_M)
\end{array}
\right],
$$

where

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

and $\mathbf{U}_m$ is the row vector of true solution states 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 the residual is characterized by

$$
\mathbf{G}\mathbf{w} - \mathbf{b} - (\mathbf{r}_0 + \mathbf{e}_{\text{int}}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{L}_{\mathbf{w}}(\mathbf{L}_{\mathbf{w}})^T). \tag{10}
$$

In the case where $\mathbf{w} = \mathbf{w}^\star$ and the integration error is negligible, (10) simplifies to

$$
\mathbf{G}\mathbf{w}^\star - \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{L}_{\mathbf{w}^\star}(\mathbf{L}_{\mathbf{w}^\star})^T). \tag{11}
$$

We note that in (11) (and in (10)), the covariance is dependent upon the parameter vector $\mathbf{w}$. In the statistical inference literature, the Iteratively Reweighted Least Squares (IRLS) [19] method offers a strategy to account for a parameter-dependent covariance by iterating between solving for $\mathbf{w}$ and updating the covariance matrix $\mathbf{C}$. In Algorithm 2 we present 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.

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 α. 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 Section 2.4.

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

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

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

$$
\mathbf{S} := \widehat{\sigma}^2 ((\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T) \, \widehat{\mathbf{C}} \, (\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1})). \tag{12}
$$

This yields the variances of individual components of $\widehat{\mathbf{w}}$ along $\mathsf{diag}(\mathbf{S})$ as well as the correlations between elements of $\widehat{\mathbf{w}}$ in the off-diagonals of $\mathbf{S}$. Here $\widehat{\sigma}^2$ is an estimate of the measurement variance σ^2, which we compute by convolving

---

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 $\{\widehat{\mathbf{w}}, \widehat{\mathbf{C}}, \widehat{\sigma}, \mathbf{S}, \text{stdx}\}$

   // 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 check $\leftarrow$ true
6 **while** <u>check</u> is true **do**
7 $\mathbf{L}^{(n)} \leftarrow [\text{mat}(\mathbf{w}^{(n)})^T \otimes \Phi]\nabla\Theta(\mathbf{U})\mathbf{P} + [\mathbb{I}_d \otimes \dot{\Phi}]$
8 $\mathbf{C}^{(n)} = (1-\alpha)\mathbf{L}^{(n)}(\mathbf{L}^{(n)})^T + \alpha\mathbf{I}$
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   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 $\widehat{\mathbf{w}} \leftarrow \mathbf{w}^{(n)}$
14 $\widehat{\mathbf{C}} \leftarrow \mathbf{C}^{(n)}$
15 $\widehat{\sigma} \leftarrow (Md)^{-1/2} \|\mathbf{f} * \mathbf{U}\|_{\text{F}}$
16 $\mathbf{S} \leftarrow \widehat{\sigma}^2((\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T)\,\widehat{\mathbf{C}}\,(\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}))$
17 stdx $\leftarrow \sqrt{\text{diag}(\mathbf{S})}$

---

each compartment of the data $\mathbf{U}$ with a high-order[6] 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 [15]), 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 σ^2 for underlying data $\mathbf{u}^\star$ that is locally well-approximated by polynomials up to degree 5.

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 [33] 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 first derive an estimator of the integration error 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 $\mathbf{m}_t := (2^\ell\underline{m}_t; \ \ell = \{0, \ldots, \bar{\ell}\})$. An SVD of this tall matrix yields an orthonormal test function

---

[6] 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 [34] Appendix F.

matrix Φ, which maximally extracts information across different scales. We note that in the later examples we have $\bar{\ell} = 3$, which in many cases leads to a largest test function support covering half of the time domain.

To begin, we consider a C^∞ bump function

$$\psi(t; a) = C \exp\left(-\frac{\eta}{[1 - (t/a)^2]_+}\right), \tag{13}$$

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

$$a = m_t \Delta t. \tag{14}$$

With the ψ in (13) we have discovered that the accuracy of the parameter estimates is relatively insensitive to a wide range of η 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 ψ to construct Φ and $\dot{\Phi}$.

Minimum radius selection

In (9), it is clear that reducing the numerical integration errors $\mathbf{e}_{\text{int}}$ will improve the estimate accuracy. However, there will be an m_t value above which the errors will be dominated by the 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 $\mathbf{e}_{\text{int}}$

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

where $\Delta t = T/M$ for a uniform timegrid $\mathbf{t} = (0, \Delta t, 2\Delta t, \ldots, M\Delta t)$ 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 $\mathbf{w}^\star$.

By expanding $\frac{d}{dt}(\phi_k(t)u^\star(t))$ into its Fourier Series[7] we then have

$$\mathbf{e}_{\text{int}}(u^\star, \phi_k, M) = \frac{T}{M\sqrt{T}}\sum_{n \in \mathbb{Z}} \mathcal{F}_n\left[\frac{d}{dt}(\phi_k(t)u^\star(t))\right]\left(\sum_{m=0}^{M-1} e^{2\pi i n m/M}\right) = \frac{2\pi i}{\sqrt{T}}\sum_{n \in \mathbb{Z}} nM\mathcal{F}_{nM}[\phi_k u^\star], \tag{15}$$

so that the integration error is entirely represented by aliased modes $\{M, 2M, \ldots\}$ of $\phi_k u^\star$. Assuming $[-a + t_k, a + t_k] \subset [0, T]$ and $T > 2a > 1$, we have the relation

$$\mathcal{F}_n[\phi_k(\cdot; a)] = a\mathcal{F}_{na}[\phi_k(\cdot; 1)],$$

hence increasing a corresponds to higher-order Fourier coefficients of $\phi_k(\cdot; 1)$ entering the error formula (15), which shows, using (15), that increasing a (eventually) lowers the integration error. For small m_t, this leads to the integration error $\mathbf{e}_{\text{int}}$ dominating the noise-related errors, while for large m_t, $\mathbf{e}_{\text{int}}$ noise-related effects are dominant.

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 $\mathbf{e}_{\text{int}}(u^\star, \phi_k, M)$ by substituting $\mathbf{U}$ for $u^\star$ and using the discrete Fourier transform (DFT), however the

---

[7] We define the nth Fourier mode of a function $f : [0, T] \to \mathbb{C}$ as $\mathcal{F}_n[f] := \frac{1}{\sqrt{T}}\int_0^T f(t)e^{-2\pi i n/T}dt$.

highest mode[8] we have access to is $\widehat{\mathcal{F}}_{\pm M/2}[\phi\mathbf{U}]$. On the other hand, we *are* able to approximate $\mathbf{e}_{\text{int}}(u^\star, \phi_k, \lfloor M/s \rfloor)$ from $\mathbf{U}$, that is, the integration error over a *coarsened* timegrid $(0, \widetilde{\Delta t}, 2\widetilde{\Delta t}, \ldots, \lfloor M/s \rfloor \widetilde{\Delta t})$, where $\widetilde{\Delta t} = T/\lfloor M/s \rfloor$ and $s > 2$ is a chosen coarsening factor. By introducing the truncated error formula

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

we have that

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

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

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

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 Figure 1)

$$\mathbf{e}_{\text{int}}(u^\star, \phi_k, M) \ \leq \ \widehat{\mathbf{e}}_{\text{int}}(\mathbf{U}, \phi_k, \lfloor M/s \rfloor, s) \ \leq \ \mathbf{e}_{\text{int}}(u^\star, \phi_k, \lfloor M/s \rfloor) \tag{16}$$

so that ensuring $\widehat{\mathbf{e}}_{\text{int}}(\mathbf{U}, \phi_k, \lfloor M/s \rfloor, s)$ is below some tolerance τ 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 $\widehat{\mathbf{e}}_{\text{int}}(u^\star, \phi_k, \lfloor M/s \rfloor, s)$, i.e.,

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

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

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

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

We pick a 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,

$$\hat{\mathbf{e}}_{\text{rms}}(m_t) := K^{-1} \sum_{k=1}^{K} \sum_{i=1}^{d} \widehat{\mathbf{e}}_{\text{int}}(\mathbf{U}^{(i)}, \phi_k(\cdot; m_t), \lfloor M/s \rfloor, s)^2, \tag{17}$$

where $\mathbf{U}^{(i)}$ is the ith variable in the system. Figure 1 depicts $\widehat{\mathbf{e}}_{\text{rms}}$ as a function of support radius m_t. As can be seen, since the variance of $\widehat{\mathbf{e}}_{\text{int}}$ 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 1 demonstrates that, in practice, the minimum radius $\underline{m}_t$ lies to the right of the changepoint of the coefficient errors

$$E_2(\widehat{\mathbf{w}}) := \left\| \widehat{\mathbf{w}} - \mathbf{w}^\star \right\|_2^2 / \left\| \mathbf{w}^\star \right\|_2^2,$$

as a function of m_t.

---

[8] We define the nth discrete Fourier mode of a function f over a periodic grid $(m\Delta t)_{m=0}^{M}$ by $\widehat{\mathcal{F}}_n[f] := \frac{\Delta t}{\sqrt{M\Delta t}} \sum_{m=0}^{M-1} f(m\Delta t)e^{-2\pi inm/M}$.
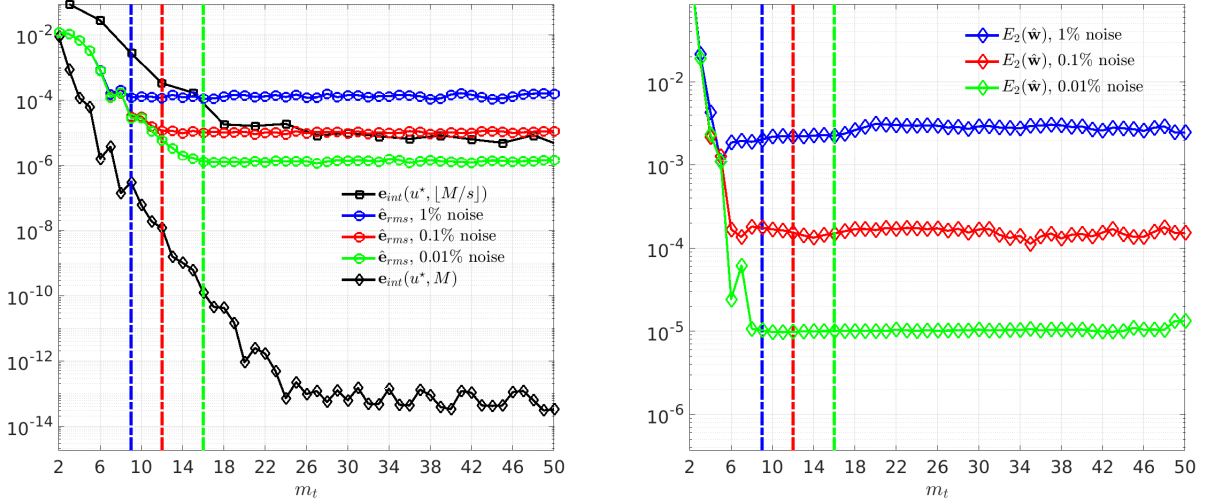
Fig. 1 Visualization of the minimum radius selection using single realizations of Fitzhugh-Nagumo data with 512 timepoints at three different noise levels. Dashed lines indicate the minimum radius $\underline{m}_t$ Left: we see that inequality (16) holds empirically for small radii m_t. Right: coefficient error E_2 as a function of m_t is plotted, showing that for each noise level the identified radius m_t using $\hat{\mathbf{e}}_{\mathrm{rms}}$ 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.

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 Ψ_ℓ be the convolution matrix for $\psi(\cdot\,; 2^\ell \underline{m}_t \Delta t)$, we compute the SVD

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

The modes $\mathbf{V}$ then form an orthonormal basis for the set of test functions forming the rows of Ψ. Letting r be the rank of Ψ, we then truncate the SVD to rank K, where K is selected as the changepoint in the cumulative sum of the singular values $(\Sigma_{ii})_{i=1}^r$. 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 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

$$
\dot{\Phi} = \mathcal{F}^{-1}\mathsf{diag}(i\boldsymbol{k})\mathcal{F}\Phi
$$

where $\mathcal{F}$ is the discrete Fourier transform and $\boldsymbol{k}$ are the requisite wavenumbers.

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 straightfoward limitations of any iterative algorithm while (3) results from the fact that our weighted least-squares framework is only approximate. In ideal scenarios where the discrepancy terms $\mathbf{e}_{\mathrm{int}}$ and $\mathbf{h}(\mathbf{u}^{\star}, \mathbf{w}^{\star}; \boldsymbol{\varepsilon})$ are negligible, equation (10) implies that

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

where $\mathbf{C}^{\star} = \mathbf{L}^{\star}(\mathbf{L}^{\star})^T$ is the covariance computed from $\mathbf{w}^{\star}$. 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 $\mathbf{w}^{\star}$. A simple way to detect this is with the Shapiro-Wilk (S-W) test for normality [51], 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 2), so we only check the S-W test after a fixed number of iterations n_0. Letting $\mathrm{SW}^{(n)} := \mathrm{SW}(\mathbf{r}^{(n)})$ denote the p-value of the S-W test at iteration $n > n_0$, and setting $\mathrm{SW}^{(n_0)} = 1$, we specify the stopping criteria as:

$$\mathrm{SC}(\mathbf{w}^{(n+1)}, \mathbf{w}^{(n)}) = \{\|\mathbf{w}^{(n+1)} - \mathbf{w}^{(n)}\|_2 / \|\mathbf{w}^{(n)}\|_2 > \tau_{\mathrm{FP}}\} \text{ and } \{n < \mathtt{max\_its}\} \text{ and } \{\mathrm{SW}^{(\max\{n, n_0\})} > \tau_{\mathrm{SW}}\}. \tag{18}$$

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

3 Illustrating Examples

Here we demonstrate the effectiveness of WENDy applied to five ordinary differential equations canonical to biology and biochemical modeling. As demonstrated in the works mentioned in Section 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 Section 3.3.

We compare the WENDy solution to the weak-form ordinary least squares solution (described in Section 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 [3, 13, 32, 39]. In most cases WENDy reduces the OLS error by 60%-90% (see the bar plots in Figures 3-7). 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.
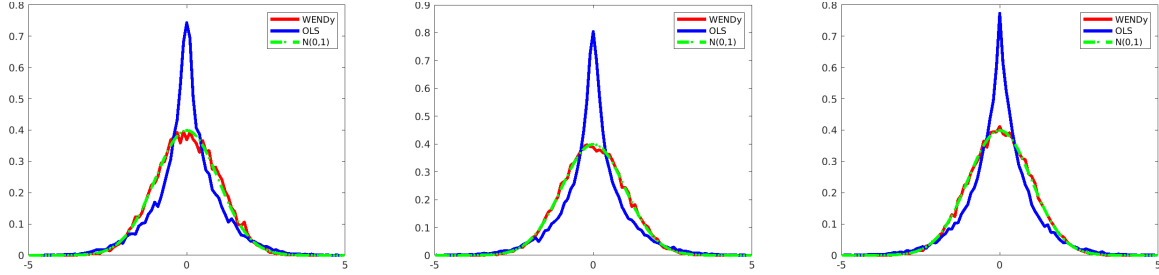
Fig. 2 Histograms of the WENDy (red) and OLS (blue) residuals evaluated at the WENDy output $\widehat{\mathbf{w}}$ applied to the (left-right) Logistic Growth, Lotka-Volterra, and Fitzhugh-Nagumo data, each with 256 timepoints and 20% noise. Curves are averaged over 100 independent trials with each histogram scaled by its empirical standard deviation. In each case, the WENDy residual agrees well with a standard normal, while the OLS residual exhibits distictly non-Gaussian features, indicative that OLS is the wrong statistical regression model.

3.1 Numerical methods and performance metrics

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

$$\sigma_{NR} \approx \frac{\|\mathsf{vec}(\mathbf{U}^\star - \mathbf{U})\|_{\mathrm{rms}}}{\|\mathsf{vec}(\mathbf{U})\|_{\mathrm{rms}}},$$

so that σ_{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 σ_{NR}, we vary the stepsize Δt (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

$$E_2 := \frac{\|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_2}{\|\mathbf{w}^\star\|_2} \tag{19}$$

as well as the forward simulation error

$$E_{\mathrm{FS}} := \frac{\|\mathsf{vec}(\mathbf{U}^\star - \widehat{\mathbf{U}})\|_2}{\|\mathsf{vec}(\mathbf{U}^\star)\|_2}. \tag{20}$$

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

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

Table 1 Specifications of ODE examples. Note that $\|\mathsf{vec}(\mathbf{U}^\star)\|_{\mathrm{rms}}$ is included for reference in order to compute the noise variance using $\sigma = \sigma_{NR}/\|\mathsf{vec}(\mathbf{U}^\star)\|_{\mathrm{rms}}$.

3.2 Summary of results

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 3 (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 (Figure 3 (top left) displays an example dataset at this resolution).

Lotka-Volterra

The Lotka-Volterra model is a system of equations designed to capture predator-prey dynamics [29]. 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 right plot in Figure 4 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).

Fitzhugh-Nagumo

The Fitzhugh-Nagumo equations are a simplified model for an excitable neuron [14]. 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, Figure 5 (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
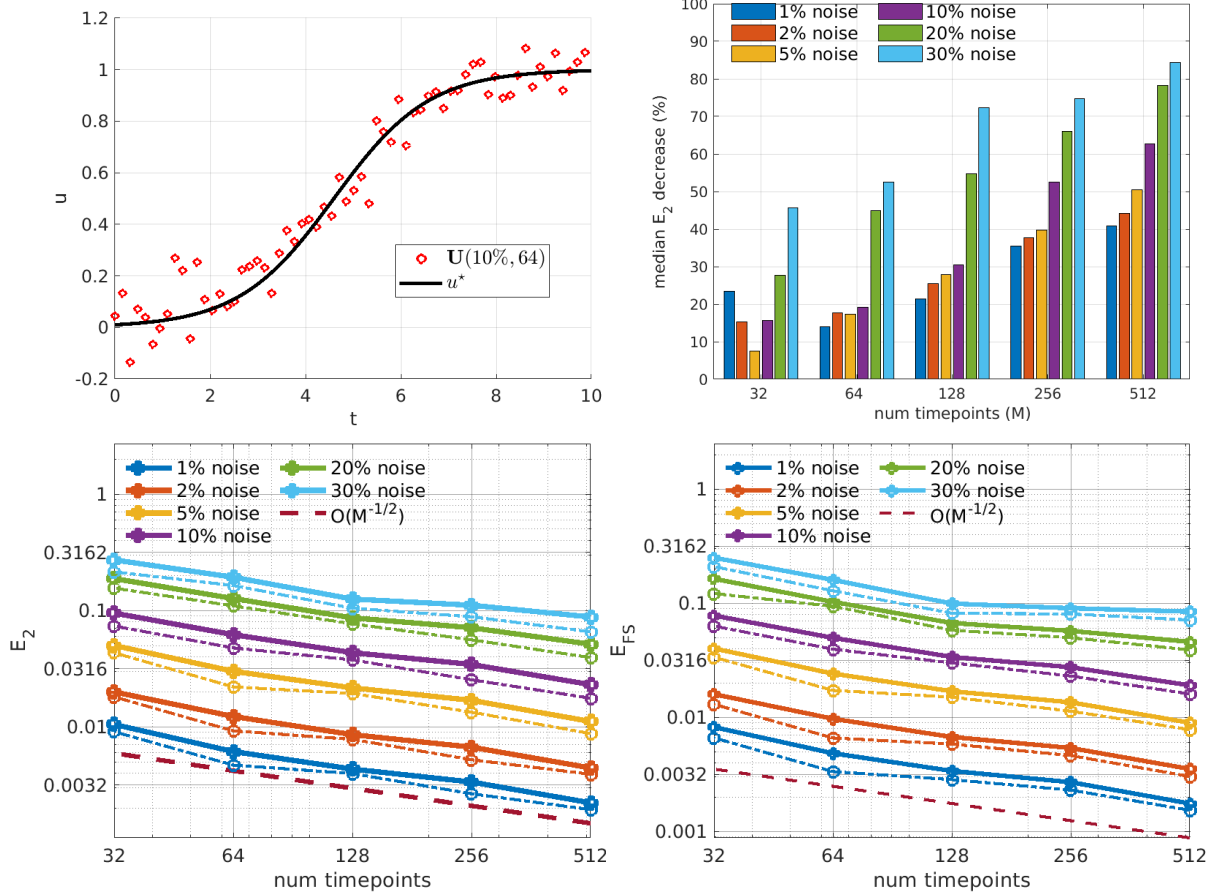
Fig. 3 Logistic Growth: Estimation of parameters in the Logistic Growth model. Left and middle panels display parameter errors E_2 and forward simulation error E_{FS}, with solid lines showing mean error and dashed lines showing median error. Right: median percentage drop in E_2 from the OLS solution to the WENDy output (e.g. at 30% noise and 512 timepoints WENDy results in a 85% reduction in error).

error by over 50% from the FSNLS solution, with 80% reductions for high noise and $M = 1024$ timepoints (top right panel). For sparse data (e.g. 64 timepoints), numerical integration errors prevent estimation of parameters with lower than 3% error, as the solution is nearly discontinuous in this case (jumps between datapoints are $\mathcal{O}(1)$).

Hindmarsh-Rose

The Hindmarsh-Rose model is used to emulate neuronal bursting and features 10 fundamental parameters which span 4 orders of magnitude [18]. 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 Figure 6, 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 6 (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%). In the operable range of 1%-2% noise and
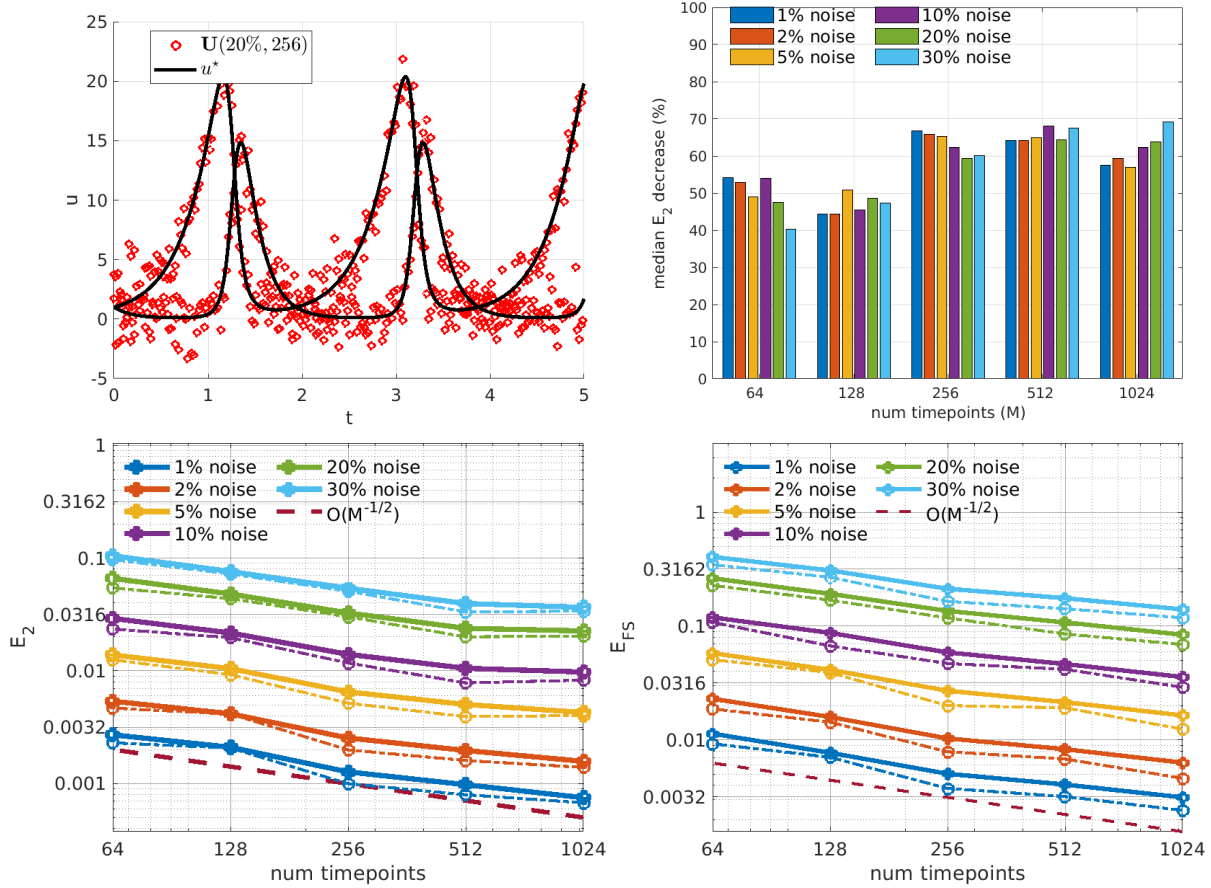
Fig. 4 Lotka-Volterra: Estimation of parameters in the Lotka-Volterra model (for plot details see Figure 3 capture).

$M \geq 256$, 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.

Protein Transduction Benchmark (PTB)

The PTB model is a five-compartment protein transduction model identified in [50] as a mechanism in the signaling cascade of epidermal growth factor (EGF). It was used in [56] to compare between four other models, and has since served as a benchmark for parameter estimation studies in biochemistry [22,30,40]. The nonlinearites 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 (Figure 7), 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.

3.3 Parameter uncertainties using learned covariance

We now demonstrate how the WENDy methodology may be used to inform the user about uncertainties in the parameter estimates. Figures 8 and 9 contain visualizations of confidence intervals around each parameter in the
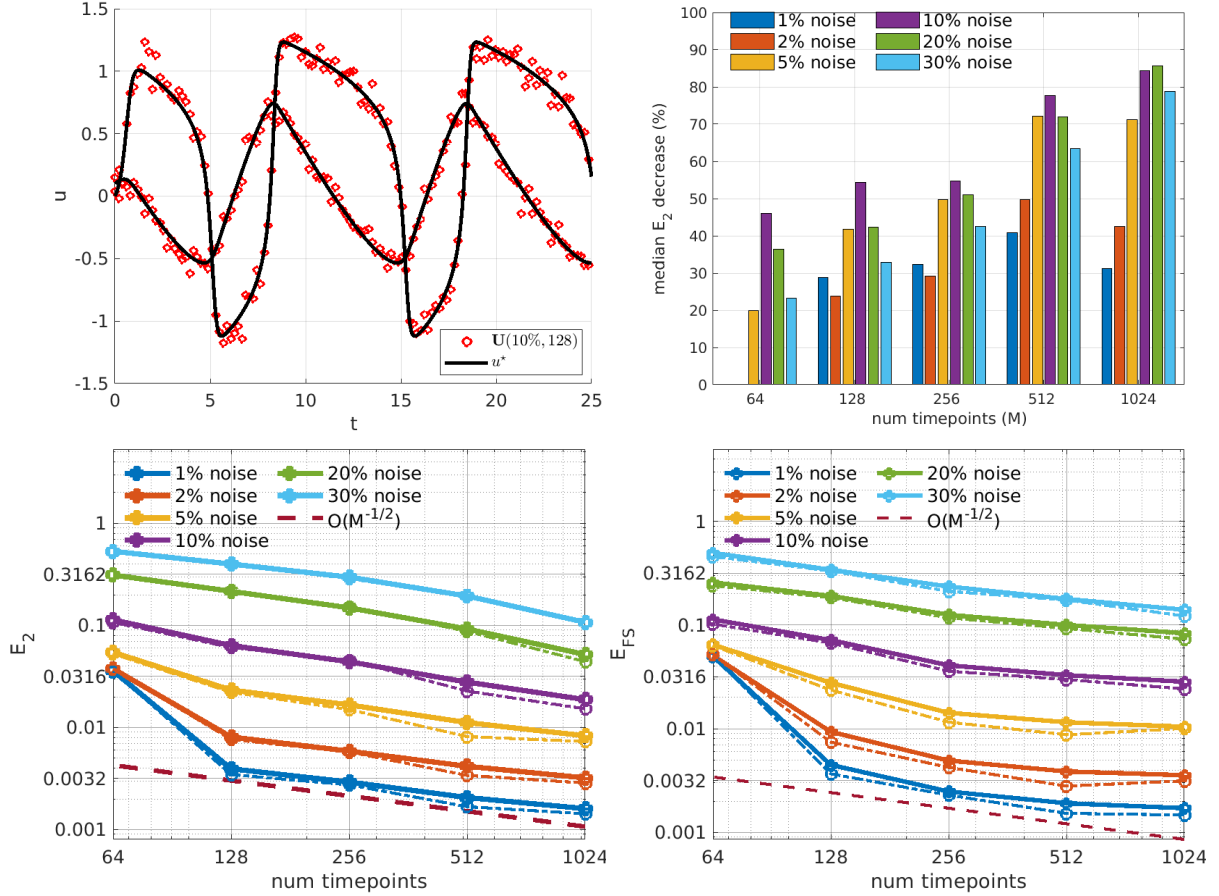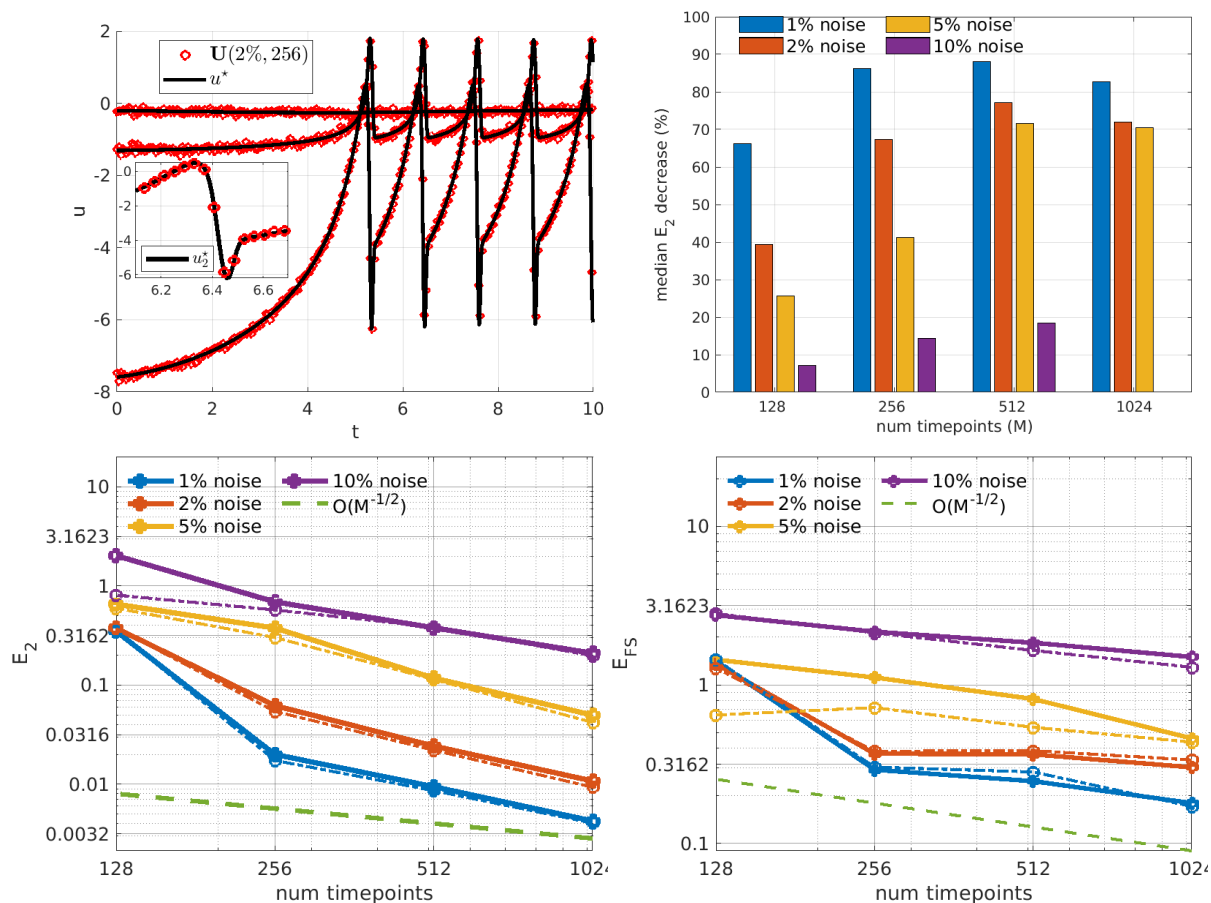
Fig. 5 FitzHugh-Nagumo: Estimation of parameters in the FitzHugh-Nagumo model (for plot details see Figure 3 capture).

FitzHugh-Nagumo and Hindmarsh-Rose models computed from the diagonal elements of the learned parameter covariance matrix **S**. Each combination of noise level and number of timepoints yields a 95% confidence interval around the learned parameter[9]. As expected, increasing the number of timepoints and decreasing the noise level leads to more certainty in the learned parameters, while lower quality data leads to higher uncertainty. Uncertainty levels can be used to inform experimental protocols and even be propagated into predictions made from learned models. One could also examine the off-diagonal correlations in **S**, which indicate how information flows between parameters. 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 2 and discussed below.

---

[9] Scripts are available at https://github.com/MathBioCU/WENDy to generate similar plots for the other examples.

Fig. 6 Hindmarsh-Rose: Estimation of parameters in the Hindmarsh-Rose model (for plot details see Figure 3 capture).

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 [38].

Due to the non-convexity of the loss function in FSNLS, choosing a good initial guess $\mathbf{w}^{(0)}$ for the parameters $\mathbf{w}^\star$ is crucial. For comparison, we use two strategies. The first strategy (simply labeled FSNLS in Figures 10-13), consists of running FSNLS on five initial guesses sampled i.i.d from a uniform distribution

$$\mathbf{w}^{(0)} \sim U(\mathbf{w}^\star, \mathbf{I}\boldsymbol{\sigma})$$

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

$$\sigma_j = 0.25|\mathbf{w}_j^\star| \tag{21}$$

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 $\mathbf{w}^\star$ are measured to high accuracy relative to the
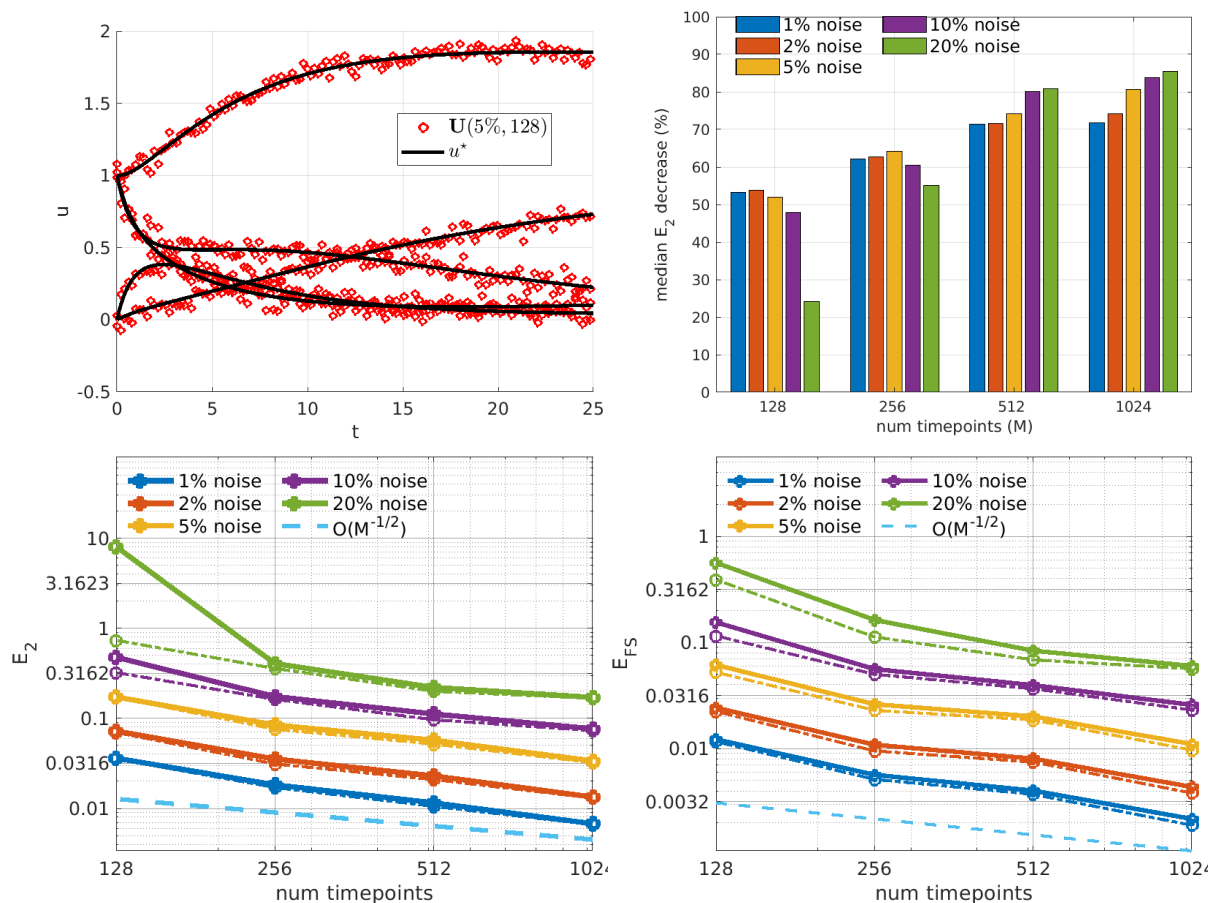
Fig. 7 Protein Transduction Benchmark (PTB): Estimation of parameters in the PTB model (for plot details see Figure 3 capture).
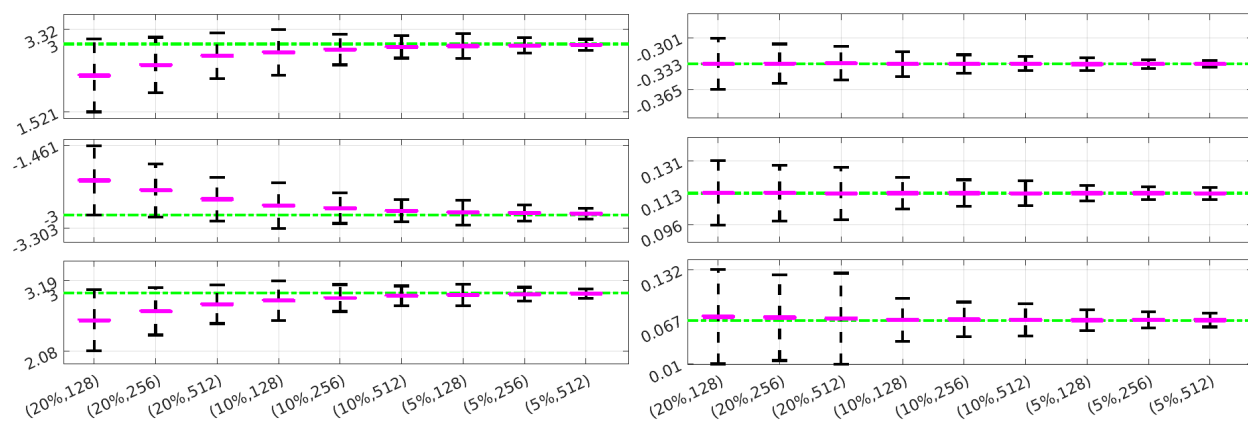
Fig. 8 FitzHugh-Nagumo: Performance of WENDy for all estimated parameters. The true parameters are plotted in green, the purple lines indicate the average learned parameters over all experiments and the black lines represent the 95% confidence intervals obtained from averaging the learned parameter covariance matrices **S**. The x-axis indicates noise level and number of timepoints for each interval.
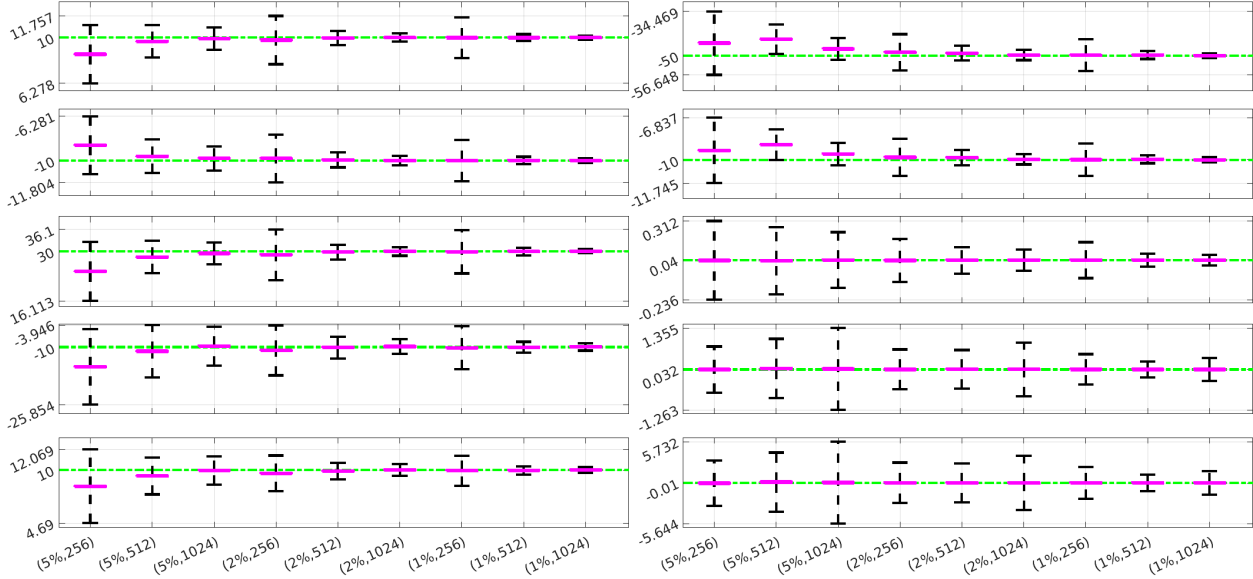
Fig. 9 Hindmarsh-Rose: Performance of WENDy for all estimated parameters. See Figure 8 for a description.

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

**Table 2** Hyperparameters for the FSNLS algorithm.

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 $\mathbf{w}^{(0)} = \widehat{\mathbf{w}}$, the output from WENDy (labeled WENDy-FSNLS in Figures 10-13). In almost all cases, this results in an increase in accuracy, and in many cases, a decrease in walltime.

Figures 10-13 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 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 Figures 10-13), all three methods are comparable in accuracy, and WENDy decreases the walltime by two orders of magnitude. In several cases, such as Lotka-Volterra Figure 10, 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 Figure 12), FSNLS is unable to provide an accurate solution ($E_2 \approx 0.2$), while WENDy and WENDy-FSNLS result in $E_2 \approx 0.005$. The clusters of FSNLS runs in Figure 12 with walltimes ≈ 10 seconds correspond to local minima, a particular weakness of FSNLS, while the remaining runs have walltimes on the order of 20 minutes, compared to 10-30 seconds WENDy. We see a similar trend in E_2 for the PTB model (Figure 13), with E_2 rarely dropping below 10%, however in this case FSNLS runs in a more reasonable amount of time, taking onlys. The WENDy solution offers speed and error reductions. For high-resolution data ($M = 1024$), WENDy runs in 40-50 seconds 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).
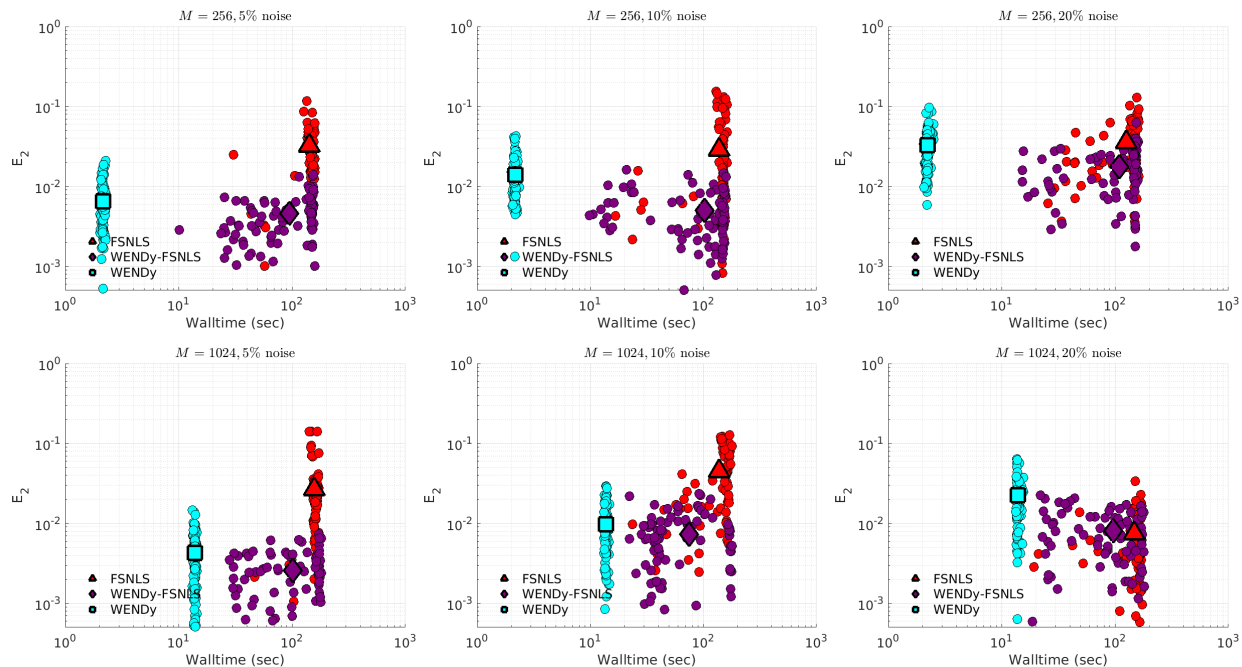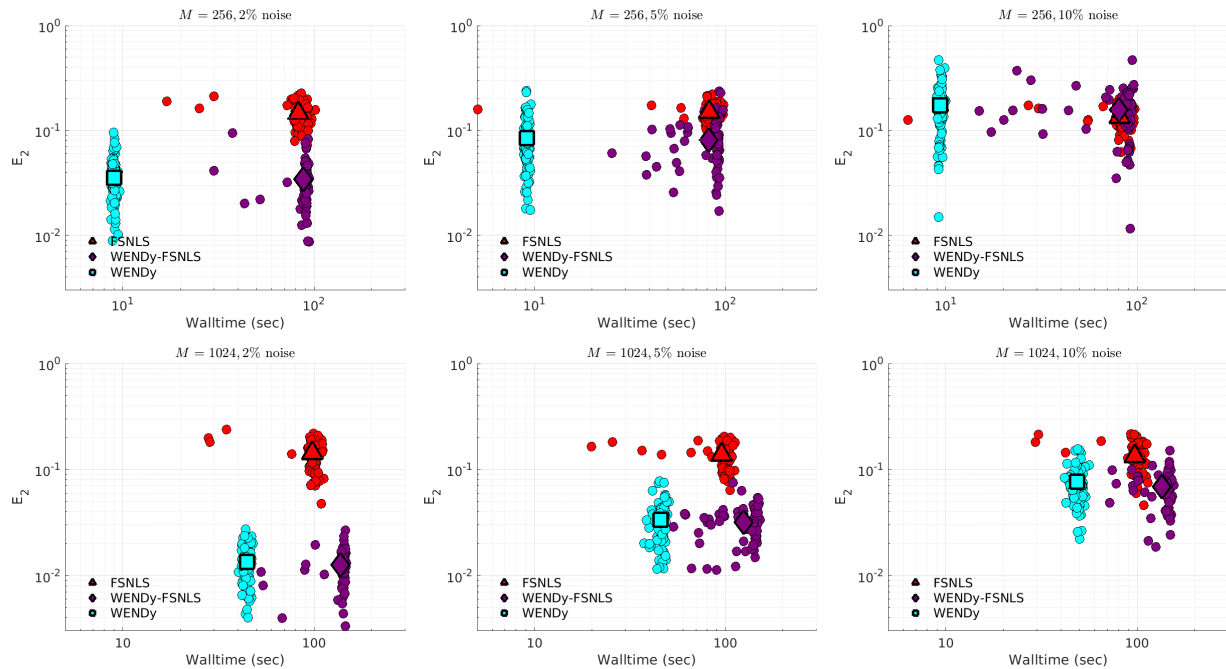
Fig. 10 Comparison between FSNLS, WENDy-FSNLS, and WENDy for the Lotka-Volterra model. Left to right: noise levels $\{5\%, 10\%, 20\%\}$. Top: 256 timepoints, bottom: 1024 timepoints. We note that the $M = 1024$ with 20% noise figure on the lower right suggests that WENDy results in slightly higher errors 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.
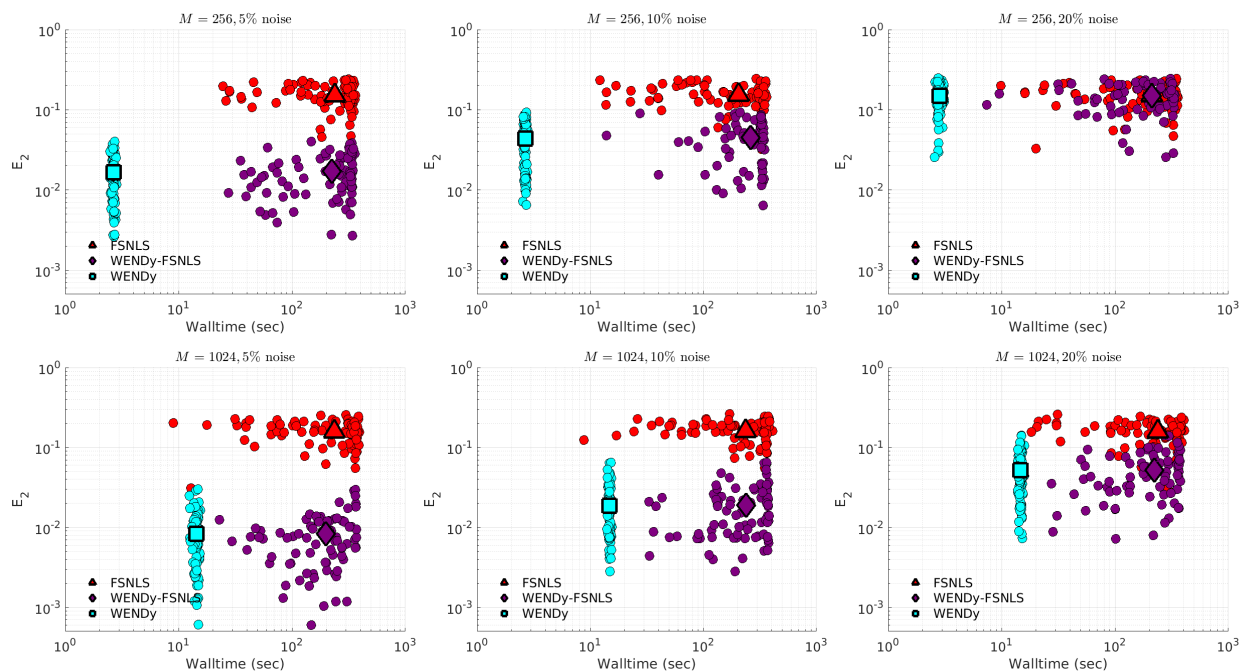
Fig. 11 Comparison between FSNLS, WENDy-FSNLS, and WENDy for the FitzHugh-Nagumo model. Left to right: noise levels $\{5\%, 10\%, 20\%\}$. Top: 256 timepoints, bottom: 1024 timepoints.
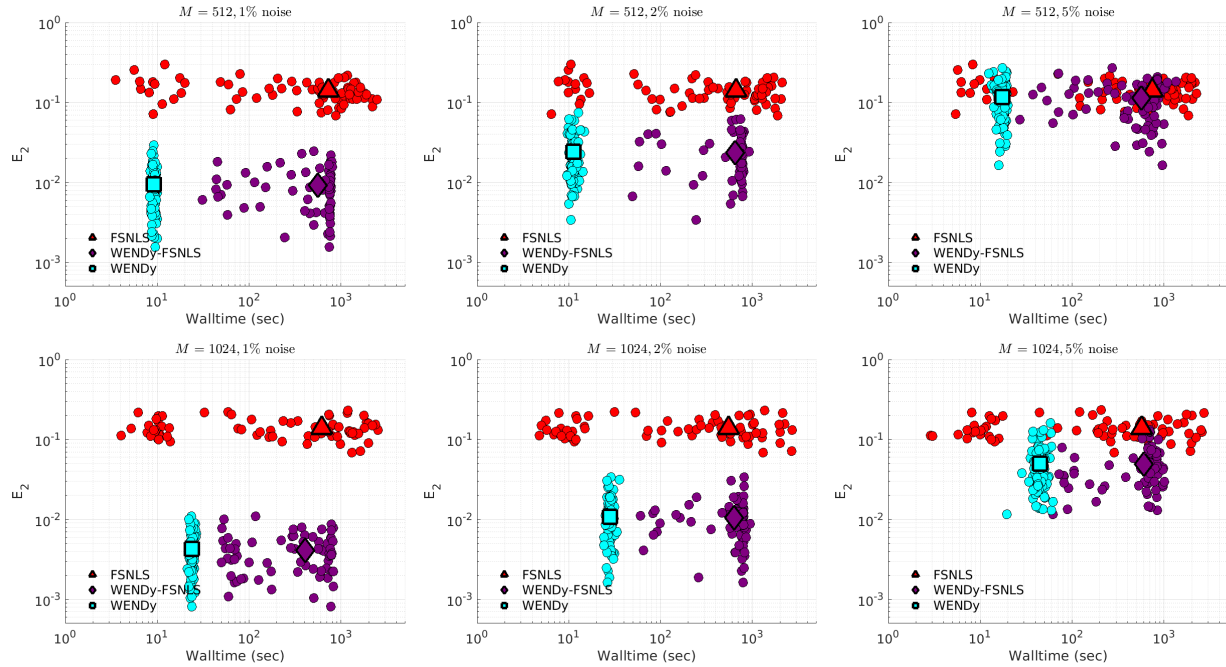
Fig. 12 Comparison between FSNLS, WENDy-FSNLS, and WENDy for the Hindmarsh-Rose model. Left to right: noise levels $\{1\%, 2\%, 5\%\}$. Top: 512 timepoints, bottom: 1024 timepoints.

Fig. 13 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.

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 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 data substitution followed by a least squares solve for the parameters has existed since at least the late 1960's [2]. However, FSNLS-based methods have proven highly successful and are ubiquitous in the parameter estimation literature and software. The disadvantage of FSNLS is that fitting using repeated forward solves comes at substantial computational cost and 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[10]. 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 $\widehat{\mathbf{C}}$.

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.

———————————

References

1. Banks, H.T., Kunisch, K.: Estimation Techniques for Distributed Parameter Systems, Systems and Control: Foundations and Applications, vol. 1. Birkhäuser Boston, Boston, MA (1989)
2. Bellman, R.: A new method for the identification of systems. Mathematical Biosciences **5**(1-2), 201–204 (1969). DOI 10.1016/0025-5564(69)90042-X
3. Bertsimas, D., Gurnee, W.: Learning sparse nonlinear dynamics via mixed-integer optimization. Nonlinear Dyn. (2023). DOI 10.1007/s11071-022-08178-9
4. Bonyadi, M.R., Michalewicz, Z.: Particle Swarm Optimization for Single Objective Continuous Space Problems: A Review. Evol. Comput. **25**(1), 1–54 (2017). DOI 10.1162/EVCO_r_00180
5. Bortz, D.M.: Accurate Model Selection Computations. J. Crit. Care **21**(4), 359 (2006)
6. Brunel, N.J.B.: Parameter estimation of ODE's via nonparametric estimators. Electron. J. Stat. **2**(0), 1242–1267 (2008). DOI 10.1214/07-EJS132
7. Brunel, N.J.B., Clairon, Q., d'Alché-Buc, F.: Parametric Estimation of Ordinary Differential Equations With Orthogonality Conditions. J. Am. Stat. Assoc. **109**(505), 173–185 (2014). DOI 10.1080/01621459.2013.841583

---

[10] This is a problem WENDy does not suffer from as there is no pre-smoothing of the data.

8. Brunton, S.L., Proctor, J.L., Kutz, J.N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc. Natl. Acad. Sci. **113**(15), 3932–3937 (2016). DOI 10.1073/pnas.1517384113

9. Calderhead, B., Girolami, M., Lawrence, N.D.: Accelerating Bayesian Inference over Nonlinear Differential Equations with Gaussian Processes. In: D. Koller, D. Schuurmans, Y. Bengio, L. Bottou (eds.) Adv. Neural Inf. Process. Syst., vol. 21. Curran Associates, Inc. (2008)

10. Dattner, I.: Differential equations in data analysis. WIREs Comp Stat **13**(6) (2021). DOI 10.1002/wics.1534

11. Dattner, I., Miller, E., Petrenko, M., Kadouri, D.E., Jurkevitch, E., Huppert, A.: Modelling and parameter inference of predator–prey dynamics in heterogeneous environments using the direct integral approach. J. R. Soc. Interface. **14**(126), 20160,525 (2017). DOI 10.1098/rsif.2016.0525

12. Ding, A.A., Wu, H.: Estimation of ordinary differential equation parameters using constrained local polynomial regression. Stat. Sin. **24**(4), 1613–1631 (2014). DOI 10.5705/ss.2012.304

13. Fasel, U., Kutz, J.N., Brunton, B.W., Brunton, S.L.: Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. ArXiv211110992 Cs Math (2021)

14. FitzHugh, R.: Impulses and Physiological States in Theoretical Models of Nerve Membrane. Biophys. J. **1**(6), 445–466 (1961). DOI 10.1016/S0006-3495(61)86902-6

15. Fornberg, B.: Generation of finite difference formulas on arbitrarily spaced grids. Math. Comput. **51**(184), 699–699 (1988). DOI 10.1090/S0025-5718-1988-0935077-0

16. Greenberg, H.: A survey of methods for determining stability parameters of an airplance from dyanmics flight measurements. Tech. Rep. NACA TN 2340, Ames Aeronautical Laboratory, Moffett Field, CA (1951)

17. Gurevich, D.R., Reinbold, P.A.K., Grigoriev, R.O.: Robust and optimal sparse regression for nonlinear PDE models. Chaos **29**(10), 103,113 (2019). DOI 10.1063/1.5120861

18. Hindmarsh, J.L., Rose, R.M.: A model of neuronal bursting using three coupled first order differential equations. Proc. R. Soc. Lond. B. **221**(1222), 87–102 (1984). DOI 10.1098/rspb.1984.0024

19. Jorgensen, M.: Iteratively Reweighted Least Squares. In: A.H. El-Shaarawi, W.W. Piegorsch (eds.) Encyclopedia of Environmetrics, first edn. Wiley (2012). DOI 10.1002/9780470057339.vai022

20. Kaptanoglu, A., de Silva, B., Fasel, U., Kaheman, K., Goldschmidt, A., Callaham, J., Delahunt, C., Nicolaou, Z., Champion, K., Loiseau, J.C., Kutz, J., Brunton, S.: PySINDy: A comprehensive Python package for robust sparse system identification. JOSS **7**(69), 3994 (2022). DOI 10.21105/joss.03994

21. Khanmohamadi, O., Xu, D.: Spatiotemporal system identification on nonperiodic domains using Chebyshev spectral operators and system reduction algorithms. Chaos **19**(3), 033,117 (2009). DOI 10.1063/1.3180843

22. Kirk, P., Thorne, T., Stumpf, M.P.: Model selection in systems and synthetic biology. Curr. Opin. Biotechnol. **24**(4), 767–774 (2013). DOI 10.1016/j.copbio.2013.03.012

23. Lax, P.D., Milgram, A.N.: IX. Parabolic Equations, <u>Annals of Mathematical Studies</u>, vol. 33, pp. 167–190. Princeton University Press (1955). DOI 10.1515/9781400882182-010

24. Liang, H., Wu, H.: Parameter Estimation for Differential Equation Models Using a Framework of Measurement Error in Regression Models. Journal of the American Statistical Association **103**(484), 1570–1583 (2008). DOI 10.1198/016214508000000797

25. Ljung, L.: System Identification: Theory for the User, second edn. Prentice Hall Information and System Sciences Series. Prentice Hall PTR, Upper Saddle River, NJ (1999)

26. Ljung, L.: System Identification, pp. 1–19. John Wiley & Sons, Inc., Hoboken, NJ, USA (2017). DOI 10.1002/047134608X.W1046.pub2

27. Loeb, J., Cahen, G.M.: Extraction a partir des enregistrements de mesures, des parametres dynamiques d'un systeme. Automatisme **8**, 479–486 (1965)

28. Loeb, J., Cahen, G.M.: More about process identification. IEEE Trans. Autom. Control **10**(3), 359–361 (1965). DOI 10.1109/TAC.1965.1098172

29. Lotka, A.J.: The Growth of Mixed Populations: Two Species Competing for a Common Food Supply, vol. 22, pp. 274–286. Springer Berlin Heidelberg, Berlin, Heidelberg (1978). DOI 10.1007/978-3-642-50151-7_12

30. Macdonald, B., Husmeier, D.: Computational Inference in Systems Biology. In: F. Ortuño, I. Rojas (eds.) Bioinformatics and Biomedical Engineering, vol. 9044, pp. 276–288. Springer International Publishing, Cham (2015). DOI 10.1007/978-3-319-16480-9_28

31. Martina-Perez, S., Simpson, M.J., Baker, R.E.: Bayesian uncertainty quantification for data-driven equation learning. Proc. R. Soc. A. **477**(2254), 20210,426 (2021). DOI 10.1098/rspa.2021.0426

32. Messenger, D.A., Bortz, D.M.: Weak SINDy For Partial Differential Equations. J. Comput. Phys. **443**, 110,525 (2021). DOI 10.1016/j.jcp.2021.110525

33. Messenger, D.A., Bortz, D.M.: Weak SINDy: Galerkin-Based Data-Driven Model Selection. Multiscale Model. Simul. **19**(3), 1474–1497 (2021). DOI 10.1137/20M1343166

34. Messenger, D.A., Bortz, D.M.: Asymptotic consistency of the WSINDy algorithm in the limit of continuum data. arXiv:2211.16000 (submitted) (2022)

35. Messenger, D.A., Bortz, D.M.: Learning mean-field equations from particle data using WSINDy. Physica D **439**, 133,406 (2022). DOI 10.1016/j.physd.2022.133406

36. Messenger, D.A., Dall'Anese, E., Bortz, D.M.: Online Weak-form Sparse Identification of Partial Differential Equations. In: Proc. Third Math. Sci. Mach. Learn. Conf., Proceedings of Machine Learning Research, vol. 190, pp. 241–256. PMLR (2022)

37. Messenger, D.A., Wheeler, G.E., Liu, X., Bortz, D.M.: Learning Anisotropic Interaction Rules from Individual Trajectories in a Heterogeneous Cellular Population. J. R. Soc. Interface **19**(195) (2022). DOI 10.1098/rsif.2022.0412

38. Nardini, J.T., Bortz, D.M.: The influence of numerical error on parameter estimation and uncertainty quantification for advective PDE models. Inverse Probl. **35**(6), 065,003 (2019). DOI 10.1088/1361-6420/ab10bb

39. Nicolaou, Z.G., Huo, G., Chen, Y., Brunton, S.L., Kutz, J.N.: Data-driven discovery and extrapolation of parameterized pattern-forming dynamics (2023)

40. Niu, M., Rogers, S., Filippone, M., Husmeier, D.: Fast Inference in Nonlinear Dynamical Systems using Gradient Matching. In: Proc. 33rd Int. Conf. Mach. Learn., vol. 48, pp. 1699–1707. PMLR (2016)

41. Pantazis, Y., Tsamardinos, I.: A unified approach for sparse dynamical system inference from temporal measurements. Bioinformatics **35**(18), 3387–3396 (2019). DOI 10.1093/bioinformatics/btz065

42. Perdreauville, F.J., Goodson, R.E.: Identification of Systems Described by Partial Differential Equations. J. Basic Eng. **88**(2), 463–468 (1966). DOI 10.1115/1.3645880

43. Poyton, A., Varziri, M., McAuley, K., McLellan, P., Ramsay, J.: Parameter estimation in continuous-time dynamic models using principal differential analysis. Computers & Chemical Engineering **30**(4), 698–708 (2006). DOI 10.1016/j.compchemeng.2005.11.008

44. Preisig, H., Rippin, D.: 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 (1993). DOI 10.1016/0098-1354(93)80001-4

45. Ramsay, J.O., Hooker, G., Campbell, D., Cao, J.: Parameter estimation for differential equations: A generalized smoothing approach. J. R. Stat. Soc. Ser. B Stat. Methodol. **69**(5), 741–796 (2007). DOI 10.1111/j.1467-9868.2007.00610.x

46. Reinbold, P.A.K., Gurevich, D.R., Grigoriev, R.O.: Using noisy or incomplete data to discover models of spatiotemporal dynamics. Phys. Rev. E **101**(1), 010,203 (2020). DOI 10.1103/PhysRevE.101.010203

47. Rudy, S.H., Brunton, S.L., Proctor, J.L., Kutz, J.N.: Data-driven discovery of partial differential equations. Sci. Adv. **3**(4), e1602,614 (2017). DOI 10.1126/sciadv.1602614

48. Sangalli, L.M.: Spatial Regression With Partial Differential Equation Regularisation. International Statistical Review **89**(3), 505–531 (2021). DOI 10.1111/insr.12444

49. Schaeffer, H., McCalla, S.G.: Sparse model selection via integral terms. Phys. Rev. E **96**(2) (2017). DOI 10.1103/PhysRevE.96.023302

50. Schoeberl, B., Eichler-Jonsson, C., Gilles, E.D., Müller, G.: Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nat Biotechnol **20**(4), 370–375 (2002). DOI 10.1038/nbt0402-370

51. Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality (complete samples). Biometrika **52**(3-4), 591–611 (1965). DOI 10.1093/biomet/52.3-4.591

52. Shinbrot, M.: On the analysis of linear and nonlinear dynamical systems for transient-response data. Tech. Rep. NACA TN 3288, Ames Aeronautical Laboratory, Moffett Field, CA (1954)

53. Van Huffel, S., Lemmerling, P. (eds.): Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications. Springer Netherlands, Dordrecht (2002). DOI 10.1007/978-94-017-3552-0

54. van Laarhoven, P.J.M., Aarts, E.H.L.: Simulated Annealing: Theory and Applications. Springer Netherlands, Dordrecht (1987). DOI 10.1007/978-94-015-7744-1

55. Varah, J.M.: A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations. SIAM J. Sci. and Stat. Comput. **3**(1), 28–46 (1982). DOI 10.1137/0903003

56. Vyshemirsky, V., Girolami, M.A.: Bayesian ranking of biochemical system models. Bioinformatics **24**(6), 833–839 (2008). DOI 10.1093/bioinformatics/btm607

57. Wang, H., Zhou, X.: Explicit estimation of derivatives from data and differential equations by Gaussian process regression. Int. J. UncertaintyQuantification **11**(4), 41–57 (2021). DOI 10.1615/Int.J.UncertaintyQuantification.2021034382

58. Wang, Z., Huan, X., Garikipati, K.: 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 (2019). DOI 10.1016/j.cma.2019.07.007

59. Wenk, P., Abbati, G., Osborne, M.A., Schölkopf, B., Krause, A., Bauer, S.: ODIN: ODE-Informed Regression for Parameter and State Inference in Time-Continuous Dynamical Systems. AAAI **34**(04), 6364–6371 (2020). DOI 10.1609/aaai.v34i04.6106

60. Xu, D., Khanmohamadi, O.: Spatiotemporal system reconstruction using Fourier spectral operators and structure selection techniques. Chaos **18**(4), 043,122 (2008). DOI 10.1063/1.3030611

61. Yang, S., Wong, S.W.K., Kou, S.C.: Inference of dynamic systems from noisy and sparse data via manifold-constrained Gaussian processes. Proc Natl Acad Sci USA **118**(15), e2020397,118 (2021). DOI 10.1073/pnas.2020397118

62. Zhang, N., Nanshan, M., Cao, J.: A Joint estimation approach to sparse additive ordinary differential equations. Stat Comput **32**(5), 69 (2022). DOI 10.1007/s11222-022-10117-y