Reconstructing Langevin systems from high and low-resolution time series using Euler and Hermite reconstructions

Open Research Statement:

Data and MATLAB codes for results reported here are available in the Github repository https://github.com/mshoja/MATLAB-reconstruction-package

Abstract. The ecological literature often features phenomenological dynamic models lacking robust validation against observational data. Reverse engineering is an alternative approach, where time series data are utilized to infer or fit a stochastic differential equation. This process, known as system reconstruction, presents significant challenges. This paper addresses the estimation of the (often) non-linear deterministic and stochastic parts of Langevin models, one of the simplest yet commonly used stochastic differential equations. We introduce a Maximum Likelihood Estimation (MLE) inference method, termed Euler reconstruction, tailored for time series data with medium to high resolution. However, the Euler approach is not reliable for low-resolution data. To fill the gap for sparsely sampled data, we present an MLE inference method pioneered by Aït-Sahalia, that we term Hermite reconstruction. We employ a powerful modeling framework utilizing splines to detect inherent nonlinearities in the unknown data-generating system to achieve high accuracy with minimal computational burden. Applying Euler and Hermite reconstructions to a range of simulated, ecological, and climate datasets, we demonstrate their efficacy and versatility. We provide a user-friendly tutorial and a MATLAB package called the 'MATLAB reconstruction package'.

Introduction. It has long been a matter of debate whether ecosystems can have alternative stable states, how to measure their resilience and whether they can recover from perturbations. These are fundamental ecological questions which have mainly been discussed using theoretical models (Connell & Sousa 1983). Tackling these questions necessitates a reverse engineering approach, where we reconstruct the system based on dynamic data (Siegert & Friedrich 2001; Rinn *et al.* 2016). This approach involves elucidating the unknown data-generating system using dynamic data. Subsequently, the resilience of the best-fitting model can be studied (Bolker *et al.* 2013; Hilborn & Mangel 2013). When appropriate mechanistic dynamical equations are known, parameters of the equations can be inferred from time series data.

Ecologists are often confronted with a situation where the nonlinear stochastic model that generated the data is unknown and must be inferred from data. If data are measured frequently throughout the entire range of the state variables, then a nonlinear stochastic model can be reconstructed by nonparametric methods (Bandi & Phillips 2003; Rinn *et al.* 2016). The reconstructed model can then be analyzed to map the stability landscape, locate stable or unstable equilibria, and calculate stochastic indicators of resilience such as mean exit time or median survival time (Arani *et al.* 2021). While nonparametric reconstruction methods are powerful, they have several limitations. They require high-resolution data and involve arbitrary choices, such as estimating the conditional mean, conditional variance, and higher conditional moments using kernels with bandwidths that control smoothness. Estimating the bandwidth is challenging, and the results may be sensitive to its choice. Additionally, these methods extrapolate moments to zero for a specified first few numbers of lags, defined as integer multiples of the time step in the data (Siegert & Friedrich 2001; Bandi & Phillips 2003; Rinn *et al.* 2016). The choice of the number of lags significantly affects the results, particularly when data resolution is low. Furthermore, these methods demand substantial amounts of data for reliable estimation (Siegert et al., 1998; Rinn et al., 2016). An alternative approach involves model-fitting techniques that maximize a likelihood function. This approach includes both the classical Euler scheme and a novel methodology pioneered by Aït-Sahalia (Aït-Sahalia 2002). It offers an attractive method that requires fewer data for reliable parameter estimation.

This paper presents a maximum likelihood estimation (MLE) approach for reconstructing the following one-dimensional Langevin models from time series data

$$dx = \mu(x;\theta)dt + \sigma(x;\theta)dW, \tag{1}$$

where $\mu(x;\theta)$ denotes the deterministic component of the system, referred to as the 'drift function' and $\sigma(x;\theta)$ represents the stochastic component, known as the 'diffusion function'. θ represents the parameter vector to be estimated via MLE and W refers to a Wiener process, where the noise source dW, known as Brownian noise, is Gaussian and white (uncorrelated). The diffusion function $\sigma(x;\theta)$ weighs the impact of perturbations, reflecting the intensity of perturbations at state x. The noise in a diffusion model (1) is additive if the diffusion function does not depend on the state x (i.e., it is constant in terms of state), otherwise, it is called multiplicative. The inference technique employed here is parametric meaning that a parametric form, with a vector of parameters θ, for the drift $\mu(x;\theta)$ and diffusion $\sigma(x;\theta)$ functions should be predefined and then the parameters are estimated using time series data through MLE.

In this study, we apply the classical inference of Langevin models using Euler methodology which we call 'Euler reconstruction'. This method requires data of medium or high resolution, posing a significant challenge in fields such as ecology and climate science. To address this issue, we present an MLE scheme

based on the groundbreaking work of Aït-Sahalia (Aït-Sahalia 2002) for univariate sparsely sampled data. The idea of Aït-Sahalia involves constructing a convergent series expansion of the conditional densities of (1) using Hermite polynomials, so we call this procedure 'Hermite reconstruction'. To our knowledge this paper is the first to provide an open-access MATLAB package and to demonstrate successful applications to ecosystem and climate data. One of our key techniques for addressing the underlying MLE problem is the utilization of a wide variety of spline models as representation of the unknown drift and diffusion functions. Splines are flexible non-linear structures in terms of the state variable x and are linear functions of the parameter vector θ, significantly enhancing the accuracy and speed of calculations (as demonstrated in various examples in our user-friendly tutorial). Additionally, splines serve as valuable tools in cases where selecting an appropriate model is uncertain, which is often encountered. To distinguish this modeling technique from typical 'parametric modeling' we term it 'spline modeling'.

The remainder of the paper presents the steps of the analysis, demonstrates reconstruction of diffusion models for several simulated, ecological, and climate data and shows that the method succeeds even for rarified datasets. Supplementary materials include a mathematical appendix, and a tutorial to illustrating a broader range of the capabilities of the Euler and Hermite reconstructions using both parametric and spline models across different data types. These data types encompass typical time series data (i.e., single datasets), replicate time series (multiple datasets belonging to the same data-generating system), big datasets, datasets with missing values, or possible combinations thereof.

Steps of the Method.

Step 1. Prepare the data: data standardization

In some real datasets, the scale of the data can be very large. In such cases, data standardization, achieved by computing z-scores (subtracting the mean and dividing by the standard deviation), makes it easier to solve the MLE. This is because standardization helps to narrow down large search spaces, making them more manageable. Additionally, standardization brings the data to a common scale, centered at 0 with small dispersion. This, in turn, makes it convenient to define a small region of parameter space for the MLE algorithm to search within (See Step4 for more details on this). For further illustration on data standardization, refer to section 8 of tutorial and Appendix G if you are interested in technical details

Step 2. Check the data requirements

There are three data requirements for a Langevin model in (1) which should be checked prior to performing the reconstruction (for a detailed discussion see section 6 of the tutorial). Firstly, data should be stationary, meaning that its statistical properties remain unchanged throughout the study period. If this assumption is

violated, the data should be divided into smaller (possibly overlapping) periods where stationarity is assured. Reconstruction can then be carried out separately for each period, with the final system reconstruction obtained by interpolating the results. Secondly, data should be Markovian meaning that the future state of data, given the present state, should be independent of the entire past history of states. The smallest time scale at which Markovicity holds is called 'Markov-Einstein' (ME) time scale (Friedrich *et al.* 2011). Reconstruction can, then, be performed on a rarified sample of data whose resolution matches the ME time scale or on samples with lower resolutions. Thirdly, it is essential to check the resolution of the rarified sample regarded now as our dataset to be analyzed. This is the subject of next step.

Step 3. Check the resolution of data: how high should the data resolution be?

How high should the data resolution be in order to be able to reconstruct the data-generating system? This is a question we should investigate prior to performing any reconstruction procedure. The answer depends on the 'speed' or 'time scale' of the yet unknown system relative to sampling frequency of the data. To investigate the time scale of data the autocorrelation of data should be examined and a quantity known as '*relaxation time*' τ_R should be estimated. For univariate data, we can roughly estimate relaxation time directly from data by fitting the exponential $\exp(-ct)$ to the initial lags of the data autocorrelation function, obtaining the estimate $\hat{c}$, and setting the relaxation time $\tau_R = 1/\hat{c}$ (refer to Appendix A for the details). Assuming Δ to be the data sampling time we consider the following three regimes.

The first regime: Δ being much smaller than τ_R

This is the desired regime and we can safely apply simple reconstruction schemes like Euler scheme or Langevin approach. Unfortunately, many real datasets are not sampled frequently enough to fall into this regime. Furthermore, some high-resolution data do not adhere to the Markov property, which is another data requirement (see Step 3). However, a rarified sample of such data may be Markovian but this comes at the expense of reduced resolution.

The second regime: Δ is approximately the same order of magnitude as τ_R

In theory, this represents the minimum resolution at which we can recover the data-generating system. In such cases, a more accurate reconstruction procedure is necessary compared to the Euler scheme. Here, Hermite reconstruction becomes essential. Refer to section 11.2 of the tutorial, which showcases a dataset accurately reconstructed with high precision using Hermite reconstruction at a resolution matching τ_R.

The third regime: Δ being bigger than τ_R

In this regime consecutive measurements are almost independent. Therefore, the true dynamics is not reflected in such datasets and any reconstruction procedure is expected to fail.

To perform a successful reconstruction, the resolution of the dataset should fall between the first two regimes. In Section 6.3 of the tutorial, we provide a convention for categorizing data resolution into three categories: low, medium, and high. Euler reconstruction is suitable for datasets categorized as medium or high resolution. However, for datasets categorized as low resolution, we employ Hermite reconstruction.

Step 4. Guidelines for implementing parametric and spline modeling

In parametric modeling, parametric forms for the drift and diffusion functions in the Langevin model (1) should be specified prior to embarking on MLE. For instance, in the Ornstein-Uhlenbeck model the parametric forms for the drift and diffusion functions are $\mu(x) = -\mu x$ and $\sigma(x) = \sigma$, respectively where $\theta = [\mu, \sigma]$ is the vector of parameters. Similarly, in the grazing model of May (May 1977) $\mu(x) = rx\left(1 - \frac{x}{K}\right) - \frac{\gamma x^2}{x^2+a^2}$, $\sigma(x) = \sigma$ where x represents the biomass of a plant population, $\theta = [r, K, \gamma, a, \sigma]$ is the vector of parameters in which, r is the growth rate, K is the carrying capacity, γ is the maximum grazing rate, a is the efficiency of the grazer, and σ is the intensity of environmental perturbations which is assumed to be constant (so, this is an additive model). In spline modeling, no model should be specified. Instead, a rather sparse mesh across the range of data, called 'knot sequence', is defined. Knots are the x-coordinate of the points where spline interpolation goes through. The values of the spline function at the knots are the model parameters (see Figures 1-5). Splines, therefore, offer flexibility in capturing unknown nonlinearities in drift and diffusion functions, making them particularly attractive when a suitable model is uncertain. Note that spline modeling is also parametric but to distinguish its rather distinct features from 'typical' parametric models we call it that way.

After selecting a model (parametric or spline), a vector of lower bounds and a vector of upper bounds for the parameters should be specified for the optimization (MLE) to search within and find the optimal parameter values. All parameters associated with the diffusion function should be bounded in a way that ensures the diffusion function remains positive. For parametric models, the physics of the problem can provide insights to determine proper vectors of lower and upper bounds. For example, in the case of the May model, we know that all model parameters should be positive as they represent ecological quantities. For spline models selecting vectors of lower and upper bounds are convenient. Since datasets we are dealing with are stationary then for the drift parameters, we can set all lower bounds to be -L and all upper bounds to be L, where L>0. Similarly, for the diffusion parameters, we can set all lower bounds to 0 and all upper bounds to L.

If a parametric modeling is preferred and it is uncertain to choose a proper model then it is advisable to select drift and diffusion models which contain constant, linear, quadratic, and higher order terms due to Taylor series expansion. An additive version of such a modeling is as follow

$$dx = \{\alpha + Ax + Bx^2 + \ldots\}dt + \sigma dW\,, \tag{2}$$

The parameter A is particularly insightful as it is the sole determinant of stability of the deterministic part of (2). The model form in (2) is not directly suited for fitting to data. Instead, data should be standardized first, and then the model (2) could be applied. Since, this model is linear in parameters, it is convenient to transform the estimated parameters back into their original scales by replacing the state variable x with its standardization $(x - m)/s$ and multiplying the right-hand side by s where m and s are the mean and standard deviation of the data, respectively.

Step 5. Euler reconstruction

If the data resolution falls within either the high-resolution or medium-resolution category, then Euler reconstruction is applicable.

Step 6. Hermite reconstruction

If the data resolution falls in the category of low resolution, Hermite reconstruction should be employed. Here, we briefly outline the approach, but for detailed mathematical explanation, refer to Appendices F, G, H and I. For a friendly explanation refer to section 11 of the tutorial, which includes numerous case studies. In our MATLAB package, we have implemented the Hermite reconstruction based on a refinement by (Bakshi & Ju 2005) to a methodology developed by (Aït-Sahalia 2002).

Our approach involves a two-phase algorithm. In the first phase Euler reconstruction is performed, resulting in a first guess of the parameter vector called $\hat{\theta}_{Euler}$. In the second phase, the algorithm improves this first guess using Hermite reconstruction. Hermite reconstruction requires two key inputs: J and K. J represents the number of 'spatial' terms in the Hermite expansion of the likelihood function using Hermite polynomials, while K represents the number of 'temporal' terms in the Taylor expansion, in Δ, of Hermite coefficients. High J and K increase estimation accuracy at the cost of higher computation time. Typically, a small J suffices, and in all case studies in the tutorial, we have used J=3. However, as data resolution decreases, a bigger K is necessary to enhance estimation accuracy. Based on our experience, for data with low-to-medium resolution a value of $K \leq 6$ suffices while for extremely low-resolution data values of $6 < K \leq 12$ are needed. Consequently, reconstructing datasets with lower resolution becomes computationally more intensive. The primary challenge lies in dealing with an optimization problem with a partially-defined objective function in the second phase. As data resolution diminishes, the algorithm should search within smaller regions in the parameter space. To address this challenge, the algorithm first identifies a set of parameters, we call 'legitimate points' (LP), where the Hermite objective function has defined values.

Subsequently, the algorithm utilizes these LPs and searches in the vicinity of the Euler estimation $\hat{\theta}_{Euler}$ to get an improved vector of parameters $\hat{\theta}_{Hermite}$. Refer to Appendix G For details of the optimization process.

Our package not only supports commonly used cubic splines but also offers a range of other spline types. The use of 'quadratic' splines is particularly significant for achieving considerably faster speeds compared to cubic splines in Hermite reconstruction. Finally, while both parametric and spline models can be used, spline modeling is generally more convenient, faster, and leads to greater accuracies.

Examples with high-resolution data

Example 1. Here, we reconstruct a high-resolution but rather small dataset with 2000 data points (see Figure 1, left panel) generated from the OU model with drift function $\mu(x) = -\mu x$, diffusion function $\sigma(x) = \sigma$, parameter values $\mu = \sigma = 1$, and sampling time $\Delta = 0.01$. We obtain the parameter estimation $\hat{\mu} = -1.0586$ and $\hat{\sigma} = 0.9953$ for this linear parametric model. We also perform a spline reconstruction to this dataset using 8 equidistance knots (see Figure 1, right panel).

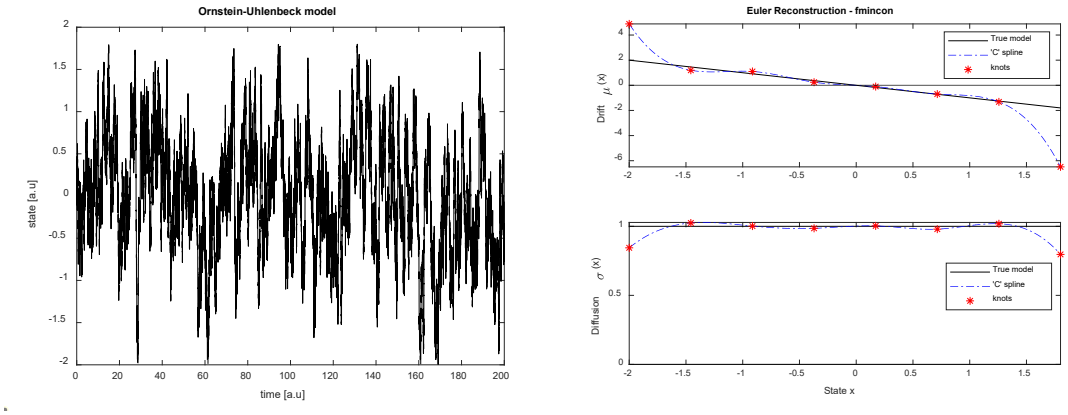

Figure 1. Illustration of spline reconstruction for a high-resolution dataset generated from a linear model. Data are illustrated in the left panel where 20000 data points, with sampling time $\Delta = 0.01$, are generated from the OU model. The right panel illustrates estimated cubic spline models (dot-dashed blue curves) for both the drift and diffusion functions using 8 regularly spaced knots over the state space alongside the true drift and diffusion functions (black curves). a.u means arbitrary unit.

Example 2. In this example we consider reconstructing a dataset (refer to Figure 2, left panel) simulated from the overgrazed model of May with drift function $\mu(x) = rx\left(1 - \frac{x}{K}\right) - \frac{\gamma x^2}{x^2 + a^2}$ and diffusion function $\sigma(x) = \sigma$ (in section 'Step4' the meaning of the parameters of May model were described). This dataset contains 10^5 data points with sampling time $\Delta = 0.01$. The parameter values are $r = 1.01, K = 10, \gamma =$

$2.75, a = 1.6, \sigma = 0.4$. Under these parameters the deterministic (i.e., without noise) model of May exhibits over-grazed and under-grazed alternative vegetation states (Figure 2, top right panel). Furthermore, the May model is nonlinear in terms of parameters K and a as well as the state variable x. Therefore, unlike the dataset in Example1, a longer dataset is needed here in order to be able to reconstruct transitions and time scales of shifts between alternative basins of attraction. Nonetheless, we could estimate parameters with a rather good accuracy as $\hat{r} = 1.1805, \widehat{K} = 9.8416, \hat{\gamma} = 3.1242, \hat{a} = 1.5279, \hat{\sigma} = 0.3997$. Figure 2, right panel illustrates a cubic spline model fitted to this dataset.

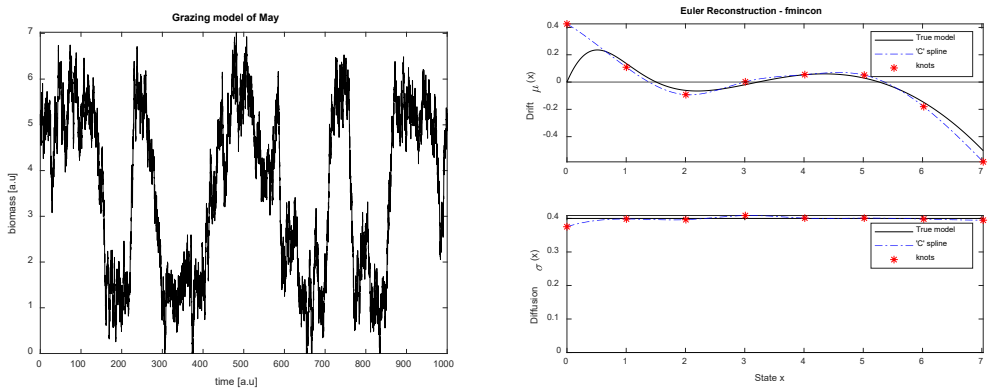

Figure 2. Illustration of spline modeling for a high-resolution dataset generated from a nonlinear model. Data are illustrated in the left panel where 10^5 data points, with sampling time $\Delta = 0.01$, are generated from the overgrazed model of May. The right panel illustrates estimated cubic spline models (dot-dashed blue curves) for both the drift and diffusion functions using 8 regularly spaced knots over the state space alongside the true drift and diffusion functions (black curves). a.u means arbitrary unit.

Example 3. In this example we reconstruct a high-resolution ecological dataset. Data is a univariate Cyanobacterial biome measured as phycocyanin concentrations in the Lake Mendota (Carpenter *et al.* 2020). The measurements were taken at minute intervals during the summer thermal stratification of 2011, a period known for common Cyanobacterial blooms (see Figure 3, left panel). For further details on this dataset, refer to references (Arani *et al.* 2021; Magnuson, Carpenter & Stanley 2023).

This dataset does not meet one of the data requirements as it lacks Markov property (indicating high correlations at the measured time scale). However, a rarified sample of this dataset, consisting of every third data point (still maintaining high resolution), does exhibit Markovian behavior. Therefore, we applied Euler reconstruction using cubic splines to the rarified sample (see Figure 3, right panel). The usefulness of spline reconstruction becomes evident here: in real datasets where choosing an appropriate model may be challenging, spline reconstruction proves to be a convenient solution.

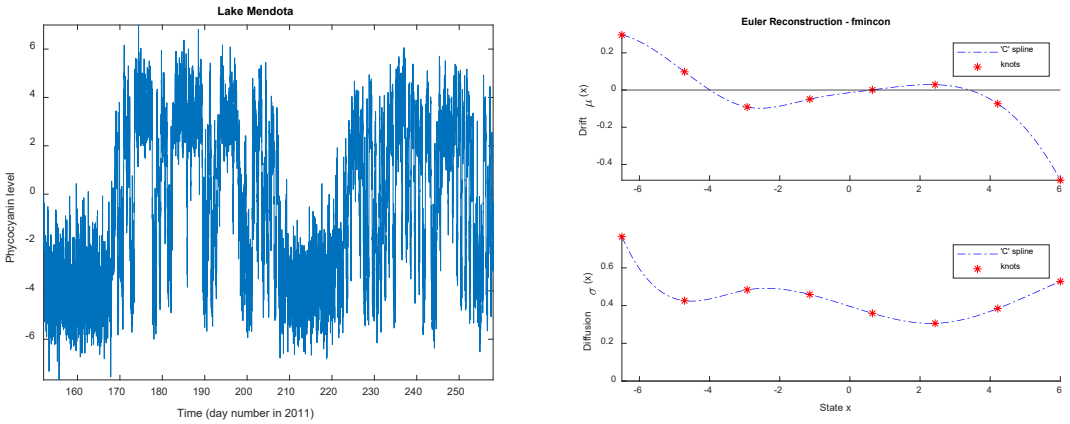

Figure 3. Illustration of spline modeling for a high-resolution ecological dataset. The left panel illustrates a high-resolution cyanobacterial measurement, taken at minute intervals, from lake Mendota. While this dataset does not meet one the data requirements (i.e., Markov property), a rarified sample of this dataset, including every third data point, satisfies this requirement. In the right panel, cubic spline modeling was applied to this rarified sample to estimate the drift and diffusion functions. Data can be found in the reference (Magnuson, Carpenter & Stanley 2023) with the URL https://doi.org/10.6073/pasta/fc8bd96677405945024ad708003be1fc

Examples with low-resolution data

Example 4. Here, we reconstruct the same dataset in Example 2 but with a resolution being 300 times less. This leads us to a low-resolution dataset. To reconstruct it we follow Hermite reconstruction which has a higher accuracy than Euler reconstruction. Here, we implement spline modeling technique. Notably, we utilize 'quadratic' splines instead of the more common cubic splines for Hermite reconstruction, significantly reducing computational time and enhancing the likelihood of successful parameter estimation. The reason has to do with the nature of Hermite reconstruction algorithm (detailed in Appendices F, G, and particularly H). Despite dataset being low-resolution, Hermite reconstruction yields remarkably accurate results in capturing the underlying data-generating system (Figure 4, right panel), outperforming Euler reconstruction (Figure 4, left panel).

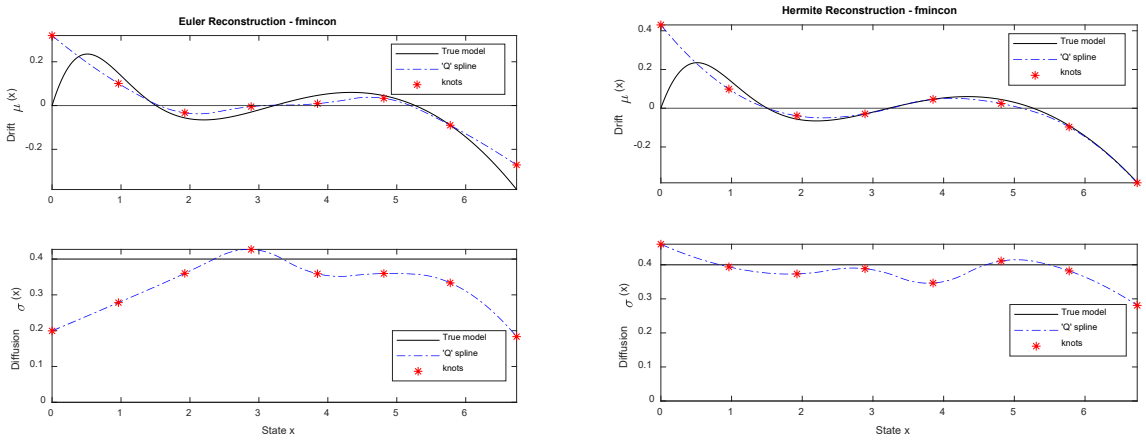

Figure 4. Illustration of Hermite and Euler reconstructions for a low-resolution dataset generated from a nonlinear model. The left panel illustrates Euler estimated quadratic spline models (dot-dashed blue curves) for both the drift and diffusion functions using 8 regularly spaced knots over the state space alongside the true drift and diffusion functions (black curves). The right panel illustrates Hermite estimated quadratic spline models (dot-dashed blue curves) for both the drift and diffusion functions using 8 regularly spaced knots over the state space alongside the true drift and diffusion functions (black curves).

Example 5. In this example, a low-resolution univariate climate dataset is reconstructed. The dataset, a $\delta^{18}O$ record from the North Greenland Ice Core Project (NGRIP) (2004), serves as a proxy for the temperature of the northern hemisphere, spanning the last 120 thousand years with a resolution of 20 years. However, this dataset fails to meet two key data requirements outlined in Step2. Initially, the dataset exhibits non-stationarity, although it stabilizes within the period from 70 to 20 thousand years before the present (see Figure 5, top panel). During this epoch, the northern hemisphere climate witnessed alternating colder (stadial) and warmer (interstadial) states, attributed to Dansgaard–Oeschger (DO) events (Dansgaard *et al.* 1993). Within the specified time frame, the majority of DO events, from DO2 to DO18 out of a total of 25 DO events, occurred (2004). Secondly, the dataset lacks the Markov property, yet a rarified sample, with every other point demonstrates Markovian behavior approximately (refer to Table 1 in section 6.3 of the tutorial for further details). Given its low resolution, Hermite reconstruction is deemed suitable. Here, we employ quadratic spline modeling to reconstruct the dataset. Figure 5 illustrates the outcomes of Euler and Hermite reconstructions. Additionally, we introduce a significant and informative quantity known as '*effective potential*' (Arani *et al.* 2021). Unlike deterministic systems where the location of equilibria can be identified using the drift function this is not the case with stochastic systems. For such systems effective potential should be used which is a quantity that incorporates information from both drift and diffusion functions. It is particularly useful for identifying alternative stable states, as is evident in this climate dataset. The minima of effective potential indicate the location of alternative stable states of stadial and interstadial

states (solid dots in Figure 6, bottom right panel) which are separated by a repellor in between (open circle in Figure 6, bottom right panel).

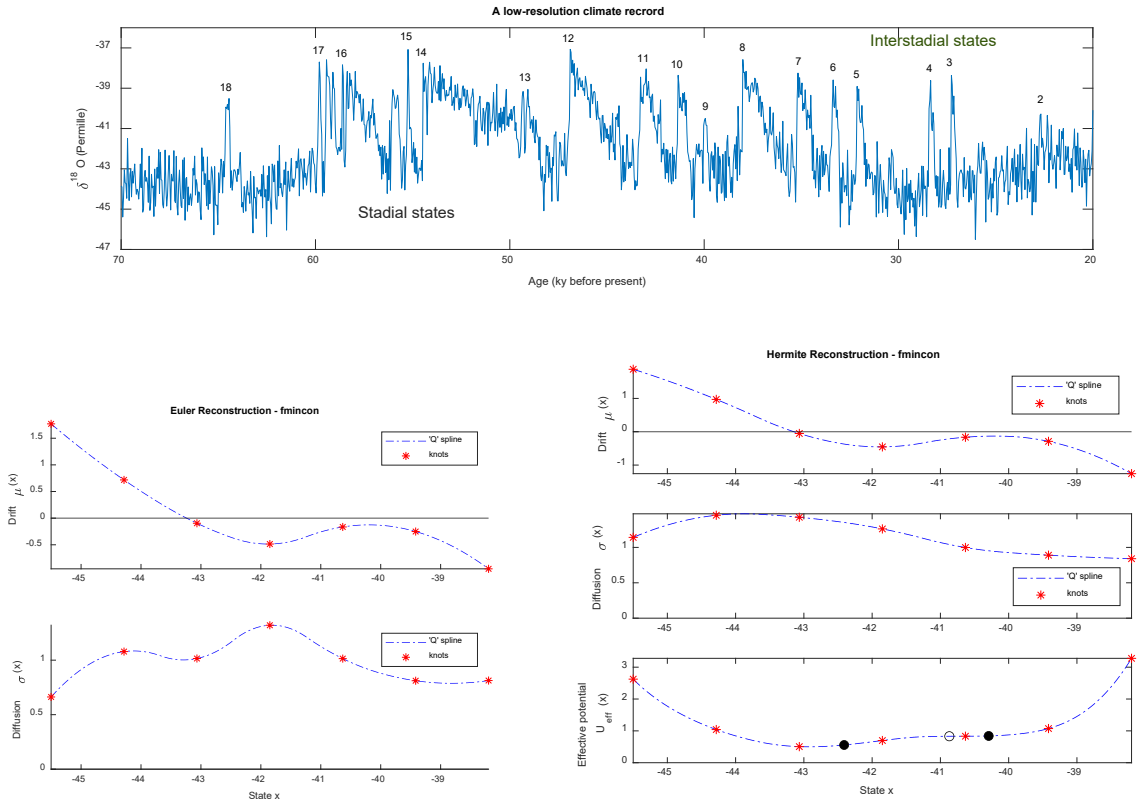

Figure 5. Illustration of Hermite and Euler reconstructions for a low-resolution climate dataset. The top panel illustrates a δ^{18}O climate record extending from 70 to 20 thousand years before the present time from NGRIP. This is used as a proxy for the temperature of the northern hemisphere which shows that the northern hemisohere climate alternated between cold stadial and warmer interstadial alternative climate states. In this time period majority of Dansgaard-Oescher events occurred (see the labels 2 to 18). The description for bottom left and right panels are similar to that in Figure 5. However, in the bottom right panel the effective potential is also depicted. Effective potential is useful to see whether there are alternative stable states in the dataset which is the case in this dataset (the solid dots represent alternative climate states of stadial and interstadial states separated by the open circle in between).

Discussion

By delineating a systematic approach encompassing data preparation, model selection, and reconstruction techniques, we furnish researchers with tools and a practical guide for analyzing diverse datasets. The methods and examples presented in this paper furnish valuable insights into the reconstruction of Langevin systems from datasets of varying resolutions. Our discussion of data requirements is fundamental for reconstructing real datasets. We acknowledge the challenges posed by datasets that do not meet these

requirements and propose strategies such as data division or rarified sampling to address these limitations. The choice between parametric and spline modeling depends on the nature of the dataset and the researcher's familiarity with the underlying dynamics, including prior empirical knowledge. While parametric models offer interpretability with respect to model form, spline models provide flexibility in capturing unknown model nonlinearities, rendering them particularly suitable when the true model is uncertain. In general, we recommend to use spline modeling. Furthermore, without incorporation of splines, addressing Hermite reconstruction for low-resolution data pose considerable challenges.

The examples presented illustrate the application of our methodology to diverse datasets, spanning from simulated data to ecological data and climate records. These examples underscore the efficacy of both Euler and Hermite reconstruction techniques, demonstrating their utility across different resolutions and system complexities. Remarkably, Hermite reconstruction proves to be particularly valuable for low-resolution datasets, offering higher accuracy compared to Euler reconstruction. This is particularly important in the fields of ecology and climate sciences, as many ecological and climate datasets have low-resolution. We believe our 'MATLAB reconstruction package' together with a step-by-step and user-friendly tutorial is a highly valuable tool for ecologists and life scientists with little affinity for mathematical and statistical modeling.

Overall, our approach furnishes a systematic framework for reconstructing complex systems from observational data. While the examples provided demonstrate the efficacy of our methodology, further research is warranted to explore its applicability to other domains and datasets, especially those generated by more complex processes than Langevin models, such as diffusion-jump models or models driven by Lévy noise. Additionally, ongoing efforts to enhance computational efficiency and address computational challenges associated with multivariate Hermite reconstruction promise to advance the field further.

References

(2004) High-resolution record of Northern Hemisphere climate extending into the last interglacial period. *Nature,* **431,** 147-151.

Aït-Sahalia, Y.J.E. (2002) Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. **70,** 223-262.

Arani, B.M., Carpenter, S.R., Lahti, L., Van Nes, E.H. & Scheffer, M. (2021) Exit time as a measure of ecological resilience. *Science,* **372,** eaay4895.

Bakshi, G. & Ju, N.J.T.J.o.B. (2005) A Refinement to Aït-Sahalia's (2002)"Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-Form Approximation Approach". **78,** 2037-2052.

Bandi, F.M. & Phillips, P.C. (2003) Fully nonparametric estimation of scalar diffusion models. *Econometrica,* **71,** 241-283.

Bolker, B.M., Gardner, B., Maunder, M., Berg, C.W., Brooks, M., Comita, L., Crone, E., Cubaynes, S., Davies, T. & de Valpine, P. (2013) Strategies for fitting nonlinear ecological models in R, AD M odel B uilder, and BUGS. *Methods in Ecology and Evolution,* **4,** 501-512.

Box, G.E. & Cox, D.R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology,* **26,** 211-243.

Carpenter, S.R., Arani, B.M., Hanson, P.C., Scheffer, M., Stanley, E.H. & Van Nes, E. (2020) Stochastic dynamics of Cyanobacteria in long-term high-frequency observations of a eutrophic lake. *Limnology and Oceanography Letters,* **5,** 331-336.

Connell, J.H. & Sousa, W.P. (1983) On the evidence needed to judge ecological stability or persistence. *The American Naturalist,* **121,** 789-824.

Dansgaard, W., Johnsen, S.J., Clausen, H.B., Dahl-Jensen, D., Gundestrup, N.S., Hammer, C.U., Hvidberg, C.S., Steffensen, J.P., Sveinbjörnsdottir, A. & Jouzel, J. (1993) Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature,* **364,** 218-220.

Ditlevsen, P.D. (1999) Observation of α-stable noise induced millennial climate changes from an ice-core record. *Geophysical Research Letters,* **26,** 1441-1444.

Friedrich, R., Peinke, J., Sahimi, M. & Tabar, M.R.R.J.P.R. (2011) Approaching complexity by stochastic methods: From biological systems to turbulence. **506,** 87-162.

Hilborn, R. & Mangel, M. (2013) *The ecological detective: confronting models with data (MPB-28)*. Princeton University Press.

Magnuson, J.J., Carpenter, S.R. & Stanley, E.H. (2023) North Temperate Lakes LTER: High Frequency Data: Meteorological, Dissolved Oxygen, Chlorophyll, Phycocyanin-Lake Mendota Buoy 2006-current.

https://doi.org/10.6073/pasta/fc8bd96677405945024ad708003be1fc

May, R.M. (1977) Thresholds and breakpoints in ecosystems with a multiplicity of stable states. *Nature,* **269,** 471-477.

Rinn, P., Lind, P.G., Wächter, M. & Peinke, J.J.a.p.a. (2016) The Langevin Approach: An R Package for Modeling Markov Processes.

Scheffer, M. (2009) *Critical transitions in nature and society*. Princeton University Press.

Scheffer, M., Carpenter, S., Foley, J.A., Folke, C. & Walker, B.J.N. (2001) Catastrophic shifts in ecosystems. **413,** 591.

Siegert, S. & Friedrich, R.J.P.R.E. (2001) Modeling of nonlinear Lévy processes by data analysis. **64,** 041107.