Comparison of continuous and discrete data-driven predictive models for hypoelliptic systems of stochastic differential equations

Fei Lu*, Kevin K. Lin†, and Alexandre J. Chorin*

Color guide

- Fei's comments
- Kevin's comments

How about this title:
Data-driven predictive modeling of hypoelliptic systems: a comparison of continuous and discrete-time approaches

k: Either is fine with me

Abstract

We compare two approaches to the inference of predictive models of dynamical systems from partial observations. The first is continuous in time, where one uses the data to infer a model in the form of a stochastic differential equation which is then discretized for numerical solution. The second is discrete in time, so that the model one infers is a parametric representation of a time series that can be directly used in computation. The ~~analysis~~ comparison is performed in a special case where the observations are known to have been obtained from a hypoelliptic stochastic equation. We show that the discrete-time approach has better predictive skills, especially when the data are relatively sparse in time, and is easier to use. The broader significance of the results is discussed.

k: see note 1

Keywords: Hypoellipticity; Langevin systems; Kramers oscillator; parameter estimation; discrete partial data; NARMA.

1 Introduction

We examine the problem of inferring predictive stochastic models for a dynamical system, given partial observations of that dynamical system at a discrete sequence of times. This inference problem arises in applications ranging from molecular dynamics and economics to climate modeling (see, e.g., [GCF15, FS01] and references therein). The systems that give rise to these observations may be stochastic or deterministic (usually chaotic). This

*Department of Mathematics, University of California, Berkeley and Lawrence Berkeley National Laboratory. E-mail addresses: feilu@berkeley.edu (FL); chorin@math.berkeley.edu (AC)

†School of Mathematical Sciences, University of Arizona. E-mail address: klin@math.arizona.edu

inference process, often called "stochastic parametrization," is useful both for reducing computational cost by constructing effective lower-dimensional models, and for making prediction possible when fully-resolved measurements of initial data and/or a full model are not available.

k: Awkward, but hopefully clear.

Stochastic parametrization often leads to hypoelliptic systems [PSW09, MH13]. Typical approaches to this problem begin by identifying a continuous-time model, usually in the form of a stochastic differential equation, then discretizing the model to make predictions. Our goal here is to compare this continuous-time approach with a fully discrete-time approach, in which one considers a discrete-time parametric model, such as a nonlinear autoregression moving average model (NARMA), and infer its parameters from the data. As we will show, advantages of the discrete-time approach include (i) it circumvents the challenging problem of estimating parameters of a stochastic differential equation from partial discrete observations; (ii) it potentially takes discretization errors into account (for a detailed discussion, see [CL15, LLC15]) and hence it can deal with large spacing between observations; (iii) it requires no further approximation before use. ~~The major~~ A main difficulty in the discrete approach is the derivation of structure, i.e. of the terms in the parametric form of the discrete-time system. We investigate in this paper the possibility of deriving structure from numerical schemes for solving SDEs.

k: see note 2

k: see note 3

~~Stochastic parametrization often leads to hypoelliptic systems~~ [PSW09, MH13]. We perform our comparison in a special case where the observations we have are known in advance to have been produced by a hypoelliptic system whose form is known and where only some parameters remain to be inferred. This choice leaves in abeyance the question of what to do in cases where much less is known about the origin of the data; in general, there is no reason to believe that a given set of observations can be described by a differential equation or by any Markovian model. We have made elsewhere [CL15, LLC15] the case that the greater generality of discrete models gives them pride of place in these more general cases. We hope that a comparison between these approaches in a relatively simple and realtively well-understood context will clarify the the advantages and disadvantages of discrete-time modeling for dynamical systems.

F: moved to previous para

k: Not sure what this means...

Model formulation and main findings. The specific hypoelliptic stochastic differential equation (SDE) we use in this paper has the form

k: see note 4

$$\begin{aligned} dx_t &= y_t\, dt, \\ dy_t &= \big(- \quad y_t - V'(x_t)\big)\, dt + \quad dB_t, \end{aligned} \tag{1.1}$$

where B_t is a standard Wiener process. When the potential V is quadratic, i.e.,

$$V(x) = \frac{}{2}x^2 , \qquad > 0,$$

we get a linear Langevin equation. When the potential has the form

$$V(x) = \frac{}{4}x^4 - \frac{}{2}x^2 , \qquad , \qquad > 0,$$

this is the Kramers oscillator [Kra40, SGH93, AI00, Hum05]. It describes the motion of a particle in a double-well potential driven by white noise, with x_t and y_t being the position

and the velocity of the particle; > 0 is a damping constant. The white noise models the thermal fluctuations of a surrounding "heat bath", the temperature of which is connected to and via the Einstein relation $T = \frac{2}{2}$. This system is ergodic, with stationary density $p(x, y) \propto \exp\left(- \frac{2}{2} \left(\frac{1}{2} y^2 + V(x) \right) \right)$. It has multiple time scales and can be highly nonlinear, but is simple enough to permit detailed numerical study, and parameter estimation for this system is well ~~understood~~ studied. ~~We wish to predict x from the discrete observations $\{x_{n\,h}\}_{n=1}^{N}$ with a separation $h > 0$; the parameters , , and are to be determinded. The variable y is not observed, hence even when the parameters are known, the initial value of y is missing when one tries to solve the SDEs to make predictions.~~

k: See new paragraph below.

Our specific goals are

- *Short-time forecasting,* i.e., to predict x from the discrete observations $\{x_{n\,h}\}_{n=1}^{N}$ with a separation $h > 0$; the parameters , , and are to be determinded. The variable y is not observed, hence even when the parameters are known, the initial value of y is missing when one tries to solve the SDEs to make predictions.

- *Estimate stationary density.* [k: someone please elaborate.]

- *Estimate time-autocovariance functions.* [k: someone please elaborate.]

Our main finding is that the discrete-time approach makes predictions as reliably as the true system that which gave rise to the data (which is of course unknown in general), even for large observation separation h, while a continuous-time approach is only accurate when h is small even in very low-dimensional examples such as ours. This suggests that for the stochastic parametrization of hypoelliptic systems, it may be advantageous to use discrete-time models. Our work also suggests that for stochastic parametrization of high-dimensional chaotic systems, even when a good parametric family of continuous-time models is available, the associated parameters may be too hard to estimate accurately, and that a discrete-time model may be more effective. Another of our findings is that numerical schemes can, in some situations, be used to select appropriate structures for discrete-time modeling.

Paper organization. We briefly review some basic facts about hypoelliptic systems in Section 2, including the parameter estimation technique we use to implement the continuous-time approach. In Section 3, we discuss the discrete-time approach in detail. Section 4 presents the numerical results, and in the Conclusion we discuss some broader implications of our results. For the convenience of the reader, we collect a number of standard results about SDEs and their numerical solutions in the Appendices.

2 Brief review of continuous-time approach

[k: I suggest adding "review," since almost all the material in this section is due to others.]

2.1 Inference of partially observed hypoelliptic systems

Consider a stochastic differential equation of the form

$$\begin{aligned} dX &= f(X, Y)\, dt \\ dY &= a(X, Y)\, dt + b(X, Y)\, dW_t \,. \end{aligned} \tag{2.1}$$

Because the stochastic forcing term is degenerate, the second-order operator in the Fokker-Planck equation

$$\frac{\partial}{\partial t}p(x,y,t) = -\frac{\partial}{\partial x}[f(x,y)p(x,y,t)] - \frac{\partial}{\partial y}[a(x,y)p(x,y,t)] + \frac{1}{2}\frac{\partial^2}{\partial y^2}[b^2(x,y)p(x,y,t)] \quad (2.2)$$

for the time evolution of probability densities is not elliptic. This means that without any further assumptions on Eq. (2.1), the solutions of the Fokker-Planck equation, and hence the transition probability associated with the SDE, might be singular in the X direction. Hypoellipticitiy is a condition that guarantees the existence of smooth solutions for Eq. (2.2) despite this degeneracy. Roughly speaking, a system is hypoelliptic if the drift terms (i.e., the vector fields $f(x,y)$ and $a(x,y)$) help to spread the noise to all phase space directions, so that the system has a nondegenerate transition density despite the degeneracy in forcing. Technically, hypoellipticity corresponds to certain conditions involving the Lie brackets of drift and diffusion fields known as Hörmander's conditions [Nua06]; when these conditions are satisfied, the system can be shown to possess smooth transition densities.

Our interest is in systems for which only discrete observations of x are available, and use these observations to estimate the parameters in the functions f, a, b. While parameter estimation for completely observed nondegenerate systems has been widely investigated (see e.g. [Rao99, Sør12]), and there has been recent progress toward parameter estimation for partially-observed nondegenerate systems [Jen14], parameter estimation for hypoelliptic systems from discrete partial observations remains challenging. There are three main categories of methods for parameter estimation (see, e.g., the surveys [Sør04], and [Sør12]):

(i) Likelihood-type methods, where the likelihood is analytically or numerically approximated, or a likelihood-type function is constructed based on approximate equations. These methods lead to maximum likelihood estimators (MLE).

(ii) Bayesian methods, in which one combines a prior with a likelihood, and one uses the posterior mean as estimator. Bayesian methods are important when the likelihood has multiple maxima. However, in many cases, suitable priors may not be available.

k: slight rephrasing

(iii) Estimation function methods, or generalized moments methods, where estimators are found by estimating functions of parameters and observations. For example, the exact likelihood can be viewed as an estimating function, and hence one may view estimating function methods as generalization of likelihood methods. To distinguish the two, we only call a method an estimating function method when the estimating function is constructed without using an (approximate) transition density. For example, the estimating function can be constructed based on martingales or moments.

Because projections of Markov processes are typically not Markov, and the system is hypoelliptic, all the above three approaches face difficulties: the likelihood function is difficult to compute, and likelihood-type functions based on approximate equations often lead to biased estimators [Glo06, PSW09, ST12]; there are no easily calculated martingales on which to base a class of estimating functions [DS04].

There are two special cases that have been well-studied. When the system is linear, the observed process is a continuous-time autoregression process. Parameter estimation for

this case is well-studied, see e.g. the review papers [Bro01, BDY07]. When the observations constitute an integrated diffusion (that is, $f(x, y) = y$ and the Y equation is autonomous, so that X is an integral of the diffusion process Y), consistent, asymptotically normal estimators are constructed in [DS04] using prediction-based estimating functions, and in [Glo06] using a likelihood type method based on Euler approximation. However, these approaches rely on the system being linear or the unobserved process being autonomous, and are not adapted to general hypoelliptic systems.

To our knowledge, for general hypoelliptic systems with discrete partial observation, only Bayesian type methods [PSW09] and a likelihood type method [ST12] have been proposed. In [PSW09] Euler and Itô-Taylor approximations are combined in a deterministic scan Gibbs sampler alternating between parameters and missing data in the unobserved variables. The reason for combining Euler and Itô-Taylor approximation is that Euler approximation leads to underestimated MLE of diffusion but is effective for drift estimation, whereas Itô-Taylor expansion leads to unbiased MLE of diffusion but is inappropriate for drift estimation. In [ST12] explicit consistent maximum likelihood type estimators are constructed based on Euler approximation of the unobserved process, where a scaling factor $\frac{3}{2}$ is used in the likelihood type function to overcome the underestimation in the diffusion (see also in [Glo06]). However, all these methods require the observation spacing h to be small and the number of observations N to be large. For example, the estimators in [ST12] are consistent under the condition that $h \to 0$, $Nh^2 \to 0$ and $Nh \to 1$. In practice, the observation spacing $h > 0$ is fixed, and large biases have been observed when h is not sufficiently small [PSW09, ST12]. As we shall show in this paper, the bias can be so large that the prediction from the estimated system may be unreliable.

k: "Spacing" instead of "gap", for consistency w/ intro

2.2 Continuous-time stochastic parametrization

The continuous-time approach starts by proposing a parametric hypoelliptic system and estimating parameters in the system from *discrete partial observations*. In the present paper, the parametric form of the hypoelliptic system is assumed to be known. Based on the Euler scheme approximation of the second equation in the system, Samson and Thieullen [ST12] constructed the following likelihood-type function, or "contrast"[1]

$$L_N(\theta) = \sum_{n=1}^{N-3} \frac{3}{2} \frac{\left[\bar{x}_{(n+2)h} - \bar{x}_{(n+1)h} + h(\bar{x}_{nh} + V'(x_{nh}))\right]^2}{h^2} + (N-3)\log \gamma^2,$$

where $\theta = (\alpha, \beta, \gamma, \gamma^2)$ and

$$\bar{x}_n = \frac{x_{(n+1)h} - x_{nh}}{h}. \tag{2.3}$$

The estimator is the minimizer of the contrast

$$\hat{\theta}_N = \arg\min_{\theta} L_N(\theta). \tag{2.4}$$

The estimator $\hat{\theta}_N$ converges to the true parameter value $\theta = (\alpha, \beta, \gamma, \gamma^2)$ under the condition that $h \to 0$, $Nh \to 1$ and $Nh^2 \to 0$. However, if h is not small enough, the

[1] Note that a shift in time in the drift term, i.e. the time index of $\bar{x}_{nh} + V'(x_{nh})$ is nh instead of $(n+1)h$, is introduced to avoid a $\sqrt{h}$ correlation between $\bar{x}_{(n+2)h} - \bar{x}_{(n+1)h}$ and $\bar{x}_{(n+1)h} + V'(x_{(n+1)h})$. Note also that there is a weighting factor $\frac{3}{2}$ in the sum.

estimator $\hat{\ }_N$ can have a large bias (see in [ST12] and in the later sections), and the bias can be so large that the estimated system may have dynamics very different from the true system and its prediction becomes unreliable.

Remark 2.1 *In the case* $V'(x) = x$*, the Langevin system* (1.1) *is linear. The process* $\{x_t, t > 0\}$ *is a continuous-time autoregressive process of order two (denoted by* CAR(2)*), and there are various ways to estimate the parameters (see the review* [Bro14]*), e.g. the likelihood method using a state-space representation and a Kalman recursion* [Jon81]*, or methods for fitting discrete-time ARMA models* [Phi59]*. However, none of these approaches can be extended to nonlinear Langevin systems. In this section we focus on methods that work for nonlinear systems.*

Once the parameters have been estimated, one numerically solves the estimated system to make predictions. In this paper, to make predictions for time $t > Nh$ (where N is the number of observations), we use the initial condition $(x_{Nh}, \bar{x}_N)$ in solving estimated system, with $\bar{x}_N$ being ~~the best guess for~~ an estimate of y_{Nh} based on observations x. Since the system is stochastic, we use an "ensemble forecasting" method to make predictions. That is, we start a number of trajectories from the same initial condition, and evolve each member of this ensemble independently. The ensemble characterizes the possible motions of the particle conditional on the past observations, and the ensemble mean provides a specific prediction. For the purpose of short-term prediction, the estimated system can be solved with small time steps, hence a low order scheme such as the Euler scheme may be used.

k: see note 5

k: see note 6

~~However, in many practical applications, the true system is unknown, one has to validate the continuous-time model by its ability of reproducing the long-term statistics of data.~~ As mentioned in the Introduction, we are also interested in long-term statistics. For this purpose, one has to compute the ergodic limits of the estimated system. The Euler scheme may be numerical unstable when the system is not globally Lipschitz, and a better scheme such as implicit Euler (see e.g. [MSH02, Tal02, MST10]) or the quasi-symplectic integrator [MT07], is needed. In our study, the Euler scheme is numerically unstable, while the Itô-Taylor scheme of strong order 2.0 in (C.2) produces long-term statistics close to those produced by the implicit Euler scheme. We use the Itô-Taylor scheme in this paper, since it has the advantage of being explicit and was used in [PSW09].

["FORECASTING ENSEMBLE OF TRAJECTORIES" HAS TO BE EXPLAINED] Did it above.

In summary, the continuous-time approach uses the following algorithm to generate a forecasting ensemble of trajectories.

Algorithm 2.2 (Continuous-time approach) *With data* $\{x_{nh}\}_{n=1}^{N}$*,*
Step 1. Estimate the parameters using (2.4)*;*
Step 2. Select a numerical scheme for the SDE, e.g. the Itô-Taylor scheme in the appendix;
Step 3. Solve the SDE (1.1) *with estimated parameters, using small time steps* dt *and initial* $\left(x_{Nh}, \frac{x_{Nh} - x_{Nh-h}}{h}\right)$*, to generate the forecasting ensemble.*

3 The discrete-time approach

3.1 NARMA representation

In the discrete-time approach, the goal is to infer a discrete-time predictive model for x from the data. Following [CL15], we choose a discrete-time system in the form of a nonlinear autoregression moving average (NARMA) model of the following form:

$$\begin{aligned} X_n &= \sum_{j=1}^{p} a_j X_{n-j} + \sum_{k=1}^{r} b_k Q_k(X_{n-p:n-1}, \xi_{n-q:n-1}) + \sum_{j=1}^{q} c_j \xi_{n-j} + \xi_n \\ &=: \mu_n + \xi_n, \end{aligned} \tag{3.1}$$

where p is the order of the autoregression, q is the order of the moving average, and the Q_k are given nonlinear functions (see below) of $(X_{n-p:n-1}, \xi_{n-q:n-1})$. Here $\{\xi_n\}$ is a sequence of i.i.d Gaussian random variables with mean zero and variance c_0^2 (denoted by $N(0, c_0^2)$). The numbers p, q, r, as well as the coefficients a_j, b_j, and c_j are to be determined from data.

A main challenge in designing NARMA models is the choice of the functions Q_k, a process we call "structure selection" or "structure derivation". Good structure design leads to models that fit data well and have good predictive capabilities. Using too many unnecessary terms, on the other hand, can lead to overfitting and inefficiency, while too few terms can lead to underfitting. As before, we assume that a parametric family containing the true model is known, and we show that suitable structures for NARMA can be derived from numerical schemes for solving SDEs. We propose the following practical criteria for structure selection: (i) the model should be numerically stable; (ii) we select the model that makes the best predictions (in practice, the predictions can be tested using the given data.); (iii) the long-time statistics of the model should agree with those of the data. These criteria are by no means optimal, and we shall discuss them further along when we discuss the numerical experiments.

k: I commented out old text because it repeats what's said later.

Once the Q_k have been chosen, the ~~parameters~~ coefficients (a_j, b_j, c_j) are estimated from data using the following conditional likelihood method. Conditional on $\xi_1, \dots, \xi_m$, the log-likelihood of $\{X_n = x_n\}_{n=m+1}^N$ is

$$L_N(\#\,|\,\xi_1, \dots, \xi_m) = \sum_{n=m+1}^{N} \frac{(X_n - \mu_n)^2}{2c_0^2} + \frac{N-q}{2} \log c_0^2,$$

where $m = \max\{p, q\}$ and $\# = (a_j, b_j, c_j, c_0^2)$. The log-likelihood is computed as follows. Conditional on given values of $\xi_1, \dots, \xi_m$, one can compute μ_{m+1} from data $\{X_n = x_n\}_{n=1}^m$. Then the value of $\{\xi_{m+1}\}$ can be computed from (3.1). Hence the values of $\{\mu_n\}_{n=m+1}^N$ and $\{\xi_n\}_{n=m+1}^N$ can be computed recursively. The estimators of the parameters $\# = (a_j, b_j, c_j, c_0^2)$ are the minimum of the log-likelihood

$$\hat{\#}_N = \arg\min_{\#} L_N(\#\,|\,\xi_1, \dots, \xi_m).$$

If the system is ergodic, the conditional maximum likelihood estimator $\hat{\#}_N$ can be proved to be consistent (see e.g. [And70, Ham94]), which means that it converges to the

true parameter value as $N \to \infty$. Hence, if N is large, $\hat{\theta}_N$ forgets about the conditional values of $\theta_1, \ldots, \theta_m$, and in practice, we can simply set $\theta_1 = \cdots = \theta_m = 0$. Also, in practice, we initialize the optimization with $c_1 = \cdots = c_q = 0$ and with the values of (a_j, b_j) computed by least-squares.

Note that in the case $q = 0$, the estimator is the same as the nonlinear least-squares estimator. ~~Note that~~ The noise sequence $\{\varepsilon_n\}$ does not have to be Gaussian for the conditional likelihood method to work.

In summary, the discrete-time approach uses the following algorithm to a generate a forecasting ensemble (see Section 2.2).

Algorithm 3.1 (Discrete-time approach) *With data* $\{x_{nh}\}_{n=1}^{N}$,
Step 1. Find possible structures for NARMA;
Step 2. Estimate the parameters in NARMA;
Step 3. Select the structure that fits the data best;
Step 4. Use the resulting model to generate a forecasting ensemble.

k: Let's move this paragraph to the Conclusion?

The advantages of the discrete-time approach over the continuous-time approach are the following. First, the estimated discrete-time system is used directly for prediction, and there is no numerical discretization step that may introduce additional errors. Second, parameter estimation is easier in the discrete-time approach. Third, as we will show, the discretization error is accounted for in modeling, hence the discrete-time approach is more tolerant of large lags between observations than the continuous time approach. Fourth, the discrete-time approach is less sensitive to model errors than the continuous time approach, since it does not require that the data be generated by a differential equation. These benefits come at the cost of needing to design suitable structures for the NARMA model, and having to redo this every time the observation spacing h changes.

k: I don't follow pt #4.

3.2 Structure derivation for the linear Langevin equation

The main difficulty in the discrete-time approach is the derivation of the structure of the NARMA model. In this section we discuss how to derive this structure from the SDEs, first in the linear case.

For the linear Langevin equation, the discrete-time system should be linear. Hence we set $r = 0$ in (3.1) and obtain an ARMA(p, q) model. The linear Langevin equation

$$\begin{aligned} dx &= y\,dt, \\ dy &= (-\gamma y - \alpha x)\,dt + \beta\,dB_t, \end{aligned} \tag{3.2}$$

can be solved analytically. The solution x_t at discrete times satisfies (see Appendix A)

$$x_{(n+2)h} = a_1 x_{(n+1)h} + a_2 x_{nh} - a_{22} W_{n+1,1} + W_{n+2,1} + a_{12} W_{n+1,2}, \tag{3.3}$$

where $\{W_{n,i}\}$ are defined in (A.1), and

$$a_1 = \operatorname{trace}(e^{Ah}),\ a_2 = -e^{-\gamma h},\ a_{ij} = \left(e^{Ah}\right)_{ij},\ \text{for } A = \begin{pmatrix} 0 & 1 \\ -\alpha & -\gamma \end{pmatrix}. \tag{3.4}$$

The process $\{x_{nh}\}$ defined in equation (3.3) is, strictly speaking, not an ARMA process (see Section B.1 for a defintion), because $\{W_{n,1}\}_{n=1}^{1}$ and $\{W_{n,2}\}_{n=1}^{1}$ are not linearly dependent and would require at least two independent noise sequences to represent, while an ARMA process requires only one. However, as the following proposition shows, there is an ARMA process with the same distribution as the process $\{x_{nh}\}$. Since the minimum mean-square-error state predictor of a stationary Gaussian process depends only on its autocovariance function (see, e.g., [BD91, Chapter 5]), an ARMA process qual in distribution to the discrete-time Langevin equation is what we need here.

k: This was the content of Remark 3.6.

Proposition 3.2 *The ARMA$(2,1)$ process*

$$X_{n+2} = a_1 X_{n+1} + a_2 X_n + W_n + \_1 W_{n-1}, \tag{3.5}$$

where a_1, a_2 are given in (3.4) and the $\{W_n\}$ are i.i.d $N(0, \_W^2)$, is the unique process in the family of invertible ARMA processes that has the same distribution as the process $\{x_{nh}\}$. Here $\_W^2$ and $\_1$ ($|\_1| < 1$ so that the process is invertible) satisfy the equations

$$\begin{aligned} \_W^2 (1 + \_1^2 + \_1 a_1) &= \_0 - \_1 a_1 - \_2 a_2, \\ \_W^2 \_1 &= \_1 (1 - a_2) - \_0 a_1, \end{aligned}$$

where $\{\_j\}_{j=0}^{2}$ are the auto-covariances of the process $\{x_{nh}\}$ and are given in Lemma A.1.

Proof. Since the stationary process $\{x_{nh}\}$ is a centered Gaussian process, we only need to find an ARMA(p,q) process with the same auto-covariance function as $\{x_{nh}\}$. The auto-covariance function of $\{x_{nh}\}$, denoted by $\{\_n\}_{n=0}^{1}$, is given by (see Lemma A.1)

$$\_n = \_0 \times \begin{cases} \frac{1}{\_1 - \_2}(\_1 e^{\_2 nh} - \_2 e^{\_1 nh}), & \text{if } \_2^2 - 4 \neq 0; \\ e^{\_0 nh}(1 - \_0 nh), & \text{if } \_2^2 - 4 = 0, \end{cases}$$

where ($\_1$, $\_2$, or $\_0$) are the roots of the characteristic polynomial $^2 + \quad + \quad = 0$ of the matrix A in (3.4).

On the other hand, the auto-covariance function of an ARMA(p,q) process

$$X_n - \_1 X_{n-1} - \cdots - \_p X_{n-p} = W_n + \_1 W_{n-1} + \cdots + \_q W_{n-q},$$

denoted as $\{(n)\}_{n=0}^{1}$, is given by (see equation (B.4))

$$(n) = \sum_{i=1}^{k} \sum_{j=0}^{r_i - 1} \_{ij} n^j \_i^{-n}, \quad \text{for } n > \max\{p, q+1\} - p,$$

where ($\_i, i = 1, \ldots, k$) are the distinct zeros of $(z) := 1 - \_1 z - \cdots - \_p z^p$, and r_i is the multiplicity of $\_i$ (hence $\sum_{i=1}^{k} r_i = p$), and $\{\_{ij}\}$ are constants.

Since $\{\_n\}_{n=0}^{1}$ only provides two possible roots, $\_i = e^{-\_i h}$ or $\_i = e^{-\_0 h}$ for $i = 1, 2$, the order p must be that $p = 2$. From these two roots, one can compute the coefficients $\_1$ and $\_2$ in the ARMA$(2,q)$ process:

$$\_1 = \_1^{-1} + \_2^{-1} = \mathrm{trace}(e^{Ah}) = a_1, \quad \_2 = -\_1^{-1} \_2^{-1} = -e^{-\_ h} = a_2.$$

Since $\gamma_k - \alpha_1 \gamma_{k-1} - \alpha_2 \gamma_{k-2} = 0$ for any $k > 2$, we have $q \neq 1$. Since $\gamma_1 - \alpha_1 \gamma_0 - \alpha_2 \gamma_1 \neq 0$, Example B.2 indicates that $q \neq 0$. Hence $q = 1$ and the above ARMA$(2,1)$ is the unique process in the family of invertible ARMA(p,q) processes that has the same distribution as $\{x_{nh}\}$. The equations for σ_w^2 and θ_1 follow from Example B.3. ■

This proposition indicates that the discrete-time system for the linear Langevin system should be an ARMA$(2,1)$ model.

Example 3.3 *Suppose* $\Delta := \gamma^2 - 4\beta < 0$. *Then the parameters in the ARMA*$(2,1)$ *process* (3.5) *are given by* $a_1 = 2e^{-\frac{\gamma}{2}h}\cos(\frac{\sqrt{-\Delta}}{2}h)$, $a_2 = -e^{-\gamma h}$ *and*

$$\theta_1 = \frac{c - a_1 - \sqrt{(c-a_1)^2 - 4}}{2}, \qquad \sigma_w^2 = \frac{\gamma_1(1-a_2) - \gamma_0 a_1}{\theta_1}.$$

where $c = \frac{\gamma_0 - \gamma_1 a_1 - \gamma_2 a_2}{\gamma_1(1-a_2) - \gamma_0 a_1}$, *and* $\gamma_n = \frac{\sigma^2}{2\beta}\left(\cos(\frac{\sqrt{-\Delta}}{2}nh) + \frac{\gamma}{\sqrt{-\Delta}}\sin(\frac{\sqrt{-\Delta}}{2}nh)\right)$ *for* $n > 0$.

Remark 3.4 *The maximum likelihood estimators of ARMA parameters can also be computed using a state-space representation and a Kalman recursion* (*see e.g.* [BD91]). *This approach is essentially the same as the conditional likelihood method in our discrete-time approach.*

Remark 3.5 *The proposition indicates that the parameters in the linear Langevin equation can also be computed from the ARMA*$(2,1)$ *estimators, because from the proof we have* $\gamma = -\frac{\ln(-a_2)}{h} = -\gamma_1 - \gamma_2$, $\beta = \theta_1\theta_2$, *and* $\sigma^2 = 2\beta\sigma_w^2$, *where* $(\theta_i, i = 1,2)$ *satisfies that* $(e^{-\gamma_i h}, i = 1,2)$ *are the two roots of* $\phi(z) = 1 - a_1 z - a_2 z$.

[k: merged remark into discussion on ARMA being equal in distribution above.]

3.3 Structure derivation for the Kramers oscillator

For nonlinear Langevin systems, in general there is no analytical solution available. We derive structures from the numerical schemes for solving stochastic differential equations. Since the goal is to derive explicit terms in a discrete-time system, implicit schemes (in e.g. [MSH02, Tal02, MT07]) are not suitable. Here we focus on deriving structures from two explicit schemes: the Euler–Maruyama scheme and the Itô-Taylor scheme of order 2.0, see Appendix C for a brief review of these schemes. As mentioned before, we expect our approach to extend to other explicit schemes, e.g., that of [AM11].

As "warm-up", we begin with the Euler-Maruyama scheme. Applying this scheme (C.1) to the system (1.1), we find:

$$\begin{aligned} x_{n+1} &= x_n + y_n h, \\ y_{n+1} &= y_n(1 - \gamma h) - hV'(x_n) + W_{n+1}, \end{aligned}$$

where $W_n = \sigma h^{1/2}\xi_n$, with $\{\xi_n\}$ is an i.id. sequence of $N(0,1)$ random variables. Substituting the first equation into the second, we obtain a closed system for x

$$x_n = (2 - \gamma h)x_{n-1} - (1 - \gamma h)x_{n-2} - h^2 V'(x_{n-2}) + hW_{n-1},$$

Note that $V'(x) = \quad x^3 - \quad x$. This leads to the following possible structure for NARMA:
Model (M1):

$$X_n = a_1 X_{n-1} + a_2 X_{n-2} + b_1 X_{n-2}^3 + \quad _n + \sum_{j=1} c_j \quad _{n-j} + \quad . \tag{3.6}$$

Next, we derive a structure from the Itô-Taylor scheme of order 2.0. Applying the scheme (C.2) to the system (1.1), we find

$$\begin{aligned} x_{n+1} &= x_n + h(1 - \quad h)\, y_n \; - h^2 V'(x_n) + Z_{n+1}, \\ y_{n+1} &= y_n \left[1 - \quad h + \quad ^2 h^2 \; - h^2 V''(x_n)\right] \; - h(1 - \quad h)\, V'(x_n) + W_{n+1} - \quad Z_{n+1}, \end{aligned}$$

where $Z_n = \quad h^{3=2}\left(\quad _n + \frac{1}{\sqrt{3}} \quad _n \right)$, with $f \quad _n g$ being an i.id. $N(0,1)$ sequence independent of $f \quad _n g$. Substituting the first equation into the second, we obtain a closed system for x:

$$\begin{aligned} x_n = x_{n-1} &\left[2 - \quad h + \quad ^2 h^2 \; - h^2 V''(x_{n-2})\right] \; - h^2 V'(x_{n-1}) + Z_n \\ &+ \left[1 - \quad h + \quad ^2 h^2 \; - h^2 V''(x_{n-2})\right] \left(- x_{n-2} + h^2 V'(x_{n-2}) \; - Z_{n-1} \right) \\ &- h^2 (1 - \quad h)^2 V'(x_{n-2}) + h(1 - \quad h)(W_{n-1} - \quad Z_{n-1}). \end{aligned}$$

Note that W_n is of order $h^{1=2}$ and Z_n is of order $h^{3=2}$. Writing the terms in descending order, we obtain

$$\begin{aligned} x_n = &\left(2 - \quad h + \quad ^2 h^2\right) x_{n-1} \; - \left(1 - \quad h + \quad ^2 h^2\right) x_{n-2} \\ &+ Z_n \; - Z_{n-1} + h(1 - \quad h)\, W_{n-1} \; - h^2 V'(x_{n-1}) + h^2 V''(x_{n-2})(x_{n-1} - x_{n-2}) \\ &+ \quad h^3 V'(x_{n-2}) + h^2 V''(x_{n-2})\, Z_{n-1} \; - h^4 V''(x_{n-2})\, V'(x_{n-2}). \end{aligned} \tag{3.7}$$

This equation suggests that $p = 2$ and $q = 0$ or 1. The noise term $Z_n \; - Z_{n-1} + h(1 - \quad h)\, W_{n-1}$ is of order $h^{1.5}$, and involves two independent noise sequences $f \quad _n g$ and $f \quad _n g$, hence the above equation for x_n is not a NARMA model. However, it suggests possible structures for NARMA models. In comparison to model (M1), the above equation has (i) different nonlinear terms of order h^2: $h^2 V'(x_{n-1})$ and $h^2 V''(x_{n-2})(x_{n-1} - x_{n-2})$; (ii) additional nonlinear terms of orders three and larger: $h^3 V'(x_{n-2})$, $h^2 Z_{n-1} V''(x_{n-2})$, and $h^4 V''(x_{n-2})\, V'(x_{n-2})$. It is not clear which terms should be used, and one may want to include as many terms as possible. However, this often leads to overfitting if too many terms are included. ~~To see this~~Hence, we consider different structures by adding more and more terms and select the one that fits data the best.. Using the fact that $V'(x) = \quad x^3 - \quad x$, these terms lead to the following possible structures for NARMA:
Model (M2):

$$X_n = a_1 X_{n-1} + a_2 X_{n-2} + b_1 X_{n-1}^3 + \underbrace{b_2 X_{n-2}^2 (X_{n-1} - X_{n-2})}_{\{z\}} + \quad _n + \sum_{j=1} c_j \quad _{n-j} + \quad ,$$

where b_1 and b_2 are of order h^2, and $q > 0$;
Model (M3):

$$\begin{aligned} X_n = a_1 X_{n-1} + a_2 X_{n-2} + b_1 X_{n-1}^3 + &\underbrace{b_2 X_{n-2}^2 (X_{n-1} - X_{n-2})}_{\{z\}} \\ + \underbrace{b_3 X_{n-2}^3}_{\{z\}} + \quad _n + &\sum_{j=1} c_j \quad _{n-j} + \quad , \end{aligned}$$

where b_3 is of order h^3, and $q > 0$;
Model (M4):

$$X_n = a_1 X_{n-1} + a_2 X_{n-2} + b_1 X_{n-1}^3 + \underbrace{b_2 X_{n-2}^2 X_{n-1}}_{\{2\}} + \underbrace{b_3 X_{n-2}^3}_{\{2\}} + \underbrace{b_4 X_{n-3}^5}_{\{2\}}$$
$$+ \underbrace{b_5 X_{n-2}^2 X_{n-1}}_{\{2\}} + \quad_n + \sum_{j=1} c_j \quad_{n-j} + \quad ,$$

where b_4 is of order h^4, and b_5 is of order $h^{3.5}$, and $q > 1$. (For the reader's convenience, we have highlighted all higher-order terms derived from $V'(x)$.)

From model (M2)–(M4), the number of nonlinear terms increases as their order increases in the numerical scheme. Following [CL15, LLC15], we forget the coefficients derived from the numerical schemes, and estimate new coefficients from data. Since there is currently no systematic way to choose the best NARMA structure, in practice we test each possible structure in turn and select the one that fits data the best.

[I DONT UNDERSTAND HOW THIS SHOWS THAT OVERFITTING IS BAD FOR YOUR, OR WHAT YOU DID TO REDUCE THE NUMBER OF TERMS] We do not do these. We propose possible models, and select the best among them.

k: see note 7

4 Simulation study

We test the continuous-time approach and the discrete-time approach for data sets with different observation intervals h. The data are generated by solving the general Langevin equation (1.1) using Itô-Taylor scheme of order 2.0, with a small step size $dt = 1/1024$, and making observations with time intervals $h = 1/32, 1/16$, and $1/8$; the value of time step dt in the integration has been chosen by trial and error, and is sufficiently small to guarantee reasonable accuracy. For each one of the data sets, we estimate the parameters in the SDE and in the NARMA models. We then compare the estimated SDE and the NARMA model by their ability to reproduce of long-term statistics and perform short-term prediction. [THERE SHOULD BE A DISCUSSION OF GOALS WAY BEFORE THIS. dit it]

4.1 The linear Langevin equation

We first discuss numerical results in the linear case. Both approaches start by computing the estimators. The estimator $\hat{} = (\hat{}, \hat{}, \hat{})$ of the parameters $(, ,)$ of the linear Langevin equation (3.2) is given by

$$\hat{} = \arg \min_{=(, ,)} \left[\sum_{n=1}^{N} \frac{3}{2} \frac{[\bar{x}_{n+2} - \bar{x}_{n+1} + h(\bar{x}_n + x_n)]^2}{h^{2}} + (N-3) \log {}^{2} \right], \qquad (4.1)$$

where $\bar{x}_n$ is computed from data using (2.3).

Following equation (3.5), we use the ARMA(2, 1) model in the discrete-time approach:

$$X_{n+2} = a_1 X_{n+1} + a_2 X_n + W_n + \quad_1 W_{n-1},$$

We estimate the parameters a_1, a_2 $_1$ and ^2_W from data using the conditional likelihood method of Section 3.1.

Table 1: Mean and standard deviation of the estimators of the parameters (, ,) of the linear Langevin equation in the continuous-time approach, computed on 100 simulations.

Estimator	True value	h = 1=32	h = 1=16	h = 1=8
ˆ	0.5	0.7313 (0.0106)	0.9538 (0.0104)	1.3493 (0.0098)
ˆ	4	3.8917 (0.0193)	3.7540 (0.0187)	3.3984 (0.0172)
ˆ	1	0.9879 (0.0014)	0.9729 (0.0019)	0.9411 (0.0023)

Table 2: Mean and standard deviation of the estimators of the parameters $(a_1, a_2, \ _1, \ _W)$ of the ARMA$(2,1)$ model in the discrete-time approach, computed on 100 simulations. The theoretical value (denoted by T-value) of the parameters are computed from proposition 3.2.

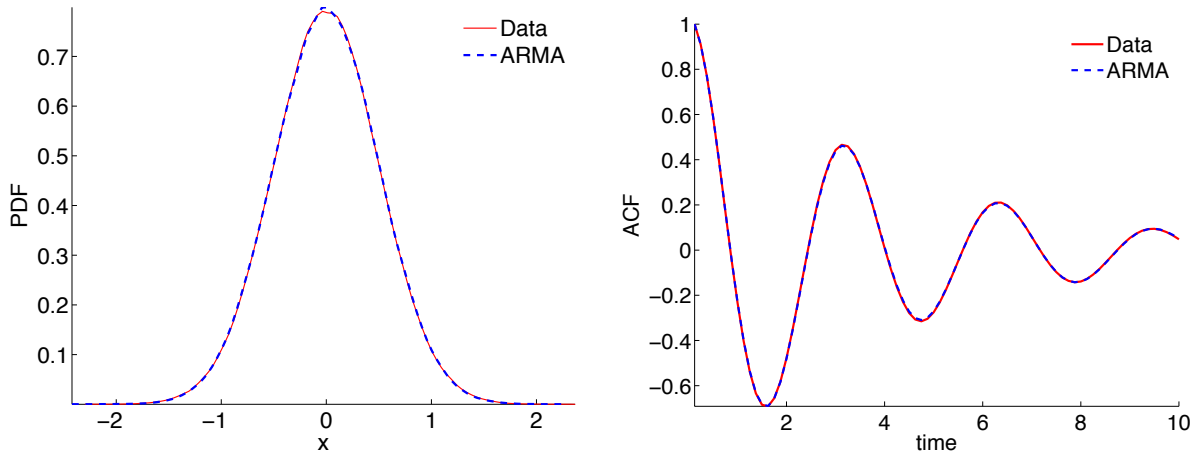
Estimator	h = 1=32		h = 1=16		h = 1=8	
	T-value	Est. value	T-value	Est. value	T-value	Est. value
$\hat{a}_1$	1.9806	1.9807 (0.0003)	1.9539	1.9541 (0.0007)	1.8791	1.8796 (0.0014)
$-\hat{a}_2$	0.9845	0.9846 (0.0003)	0.9692	0.9695 (0.0007)	0.9394	0.9399 (0.0014)
$\hat{}_1$	0.2681	0.2667 (0.0017)	0.2684	0.2680 (0.0025)	0.2698	0.2700 (0.0037)
$\hat{}_W$	0.0043	0.0043 (0.0000)	0.0121	0.0121 (0.0000)	0.0336	0.0336 (0.0001)

First, we investigate the reliability of the estimators. A hundred simulated data sets are generated from equation (3.2) with true parameters = 0.5, = 4, and = 1, and with initial $x_0 = y_0 = \frac{1}{2}$ and time interval $[0, 10^4]$. The estimators, of (, ,) in the linear Langevin equation and of $(a_1, a_2, \ _1, \ _W)$ in the ARMA$(2,1)$ model, are computed for each data-set. Empirical mean and standard deviation of the estimators are reported in Table 1 for the continuous-time approach, and Table 2 for the discrete-time approach. In the continuous-time approach, the biases of the estimators grow as h increases. In particular, large biases occur for the estimators of : the bias of ˆ increases from 0.2313 when h = 1=32 to 0.4879 when h = 1=8, while the true value is = 0.5; similarly large biases of estimators were also noticed in [ST12]. In contrast, the biases are much smaller for the discrete-time approach. ~~In the discrete-time approach, on the other hand, the biases in the estimators are small.~~ The "theoretical value" (denoted by "T-value") of $a_1, a_2, \ _1$ and ^2_W are computed analytically as in Example 3.3. Table 2 shows that the estimators in the discrete-time approach have negligible differences from the theoretical values.

In practice, the above test of the reliability of estimators cannot be performed, because one has only a single dataset and the true system that generated the data is unkown.

k: Do we even need to say this?

We now compare the two approaches in a practical setting, by assuming that we are only given a single data set from discrete observations of a long trajectory on time interval $[0, T]$ with $T = 2^{18} \approx 3 \times 10^5$. We estimate the parameters in the SDE and the ARMA model, and again investigate the performance of the estimated SDE and NARMA model in reproducing long-term statistics and in predicting the short-term evolution of x. The long-term statistics are computed by time-averaging. The first half of the data set is used to compute the estimators, and the second half of the dataset is used to test the prediction.



(a) PDF (b) ACF

Figure 1: Empirical PDF and ACF of the ARMA$(2,1)$ model in the case $h = 1/8$. The other two cases $h = 1/32$ and $h = 1/16$ are almost identical. The Langevin system with estimated parameters can not reproduce the PDF and ACF because it is numerically unstable.

The long-term statistics, i.e., the empirical probability density function (PDF) and the autocorrelation function (ACF), are shown in Figure 1. For all three observation separations h, the estimated SDEs cannot reproduce the long-term statistics, because they are numerically unstable, even when the Itô-Taylor scheme of order 2.0 with the time-step $dt = 1/1024$ (which was used to generate data) is used. In contrast, the AMRA model reproduces the empirical PDF and ACF almost perfectly for all three values of h.

Next, we use an ensemble of trajectories to predict the motion of x. For each ensemble, we calculate the mean trajectory and compare it with the true trajectory from the data. We measure the performance of the prediction by computing the root-mean-square-error (RMSE) and the anomaly correlation (ANCR) of a large number of ensembles as follows: take N_0 short pieces of data from the second half of the long trajectory, denoted by $\left(x_{(n_i+1)h}, \ldots, x_{(n_i+K)h}\right)_{i=1}^{N_0}$, where $n_i = Ki$. For each short piece of data $\left(x_{(n_i+1)h}, \ldots, x_{(n_i+K)h}\right)$, we generate N_{ens} trajectories $\left(X_1^{i,j}, \ldots, X_K^{i,j}\right)_{j=1}^{N_{ens}}$ using a prediction system (i.e., the NARMA(p,q), the estimated Langevin system, or the true Langevin system), starting all ensemble members from the same several-step initial condition $\left(x_{(n_i+1)h}, \ldots, x_{(n_i+m)h}\right)$, where $m = 2\max\{p,q\} + 1$. For the NARMA(p,q) we start with $\epsilon_1 = \cdots = \epsilon_q = 0$. For the estimated Langevin system and the true Langevin system, we start with inital $\left(x_{(n_i+m)h}, \frac{x_{(n_i+m)h} - x_{(n_i+m-1)h}}{h}\right)$ and solve them using the Itô-Taylor scheme of order 2.0 with a a small time step $dt = 1/64$ and record the trajectories every h/dt steps to get the prediction trajectories $\left(X_1^{i,j}, \ldots, X_K^{i,j}\right)$.

We then calculate the mean trajectory for each ensemble, $\overline{X}_k^i = \frac{1}{N_{ens}} \sum_{j=1}^{N_{ens}} X_k^{i,j}$, $k = 1, \ldots, K$. The RMSE measures, in an average sense, the difference between the mean ensemble trajectory and the true data trajectory:

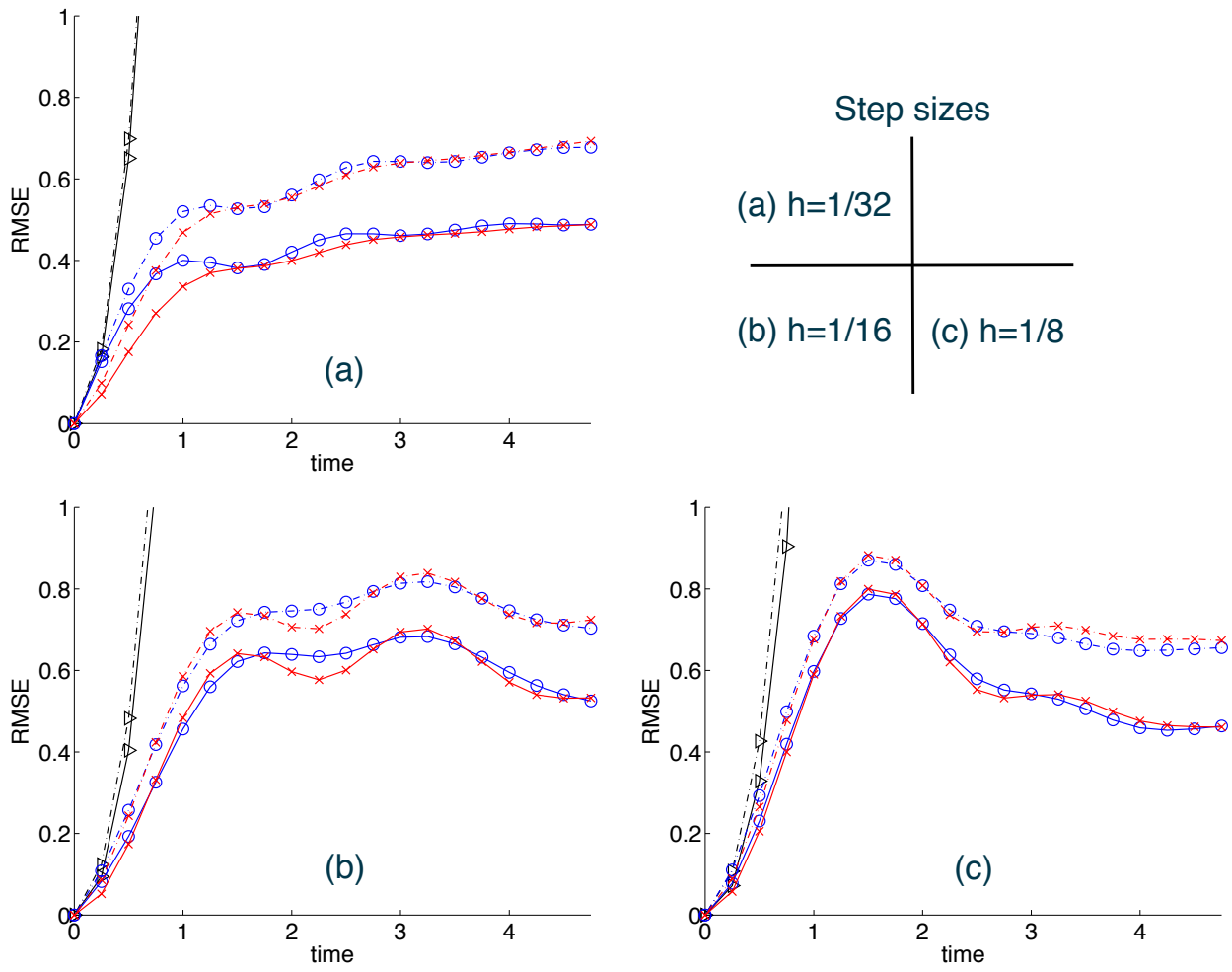


Figure 2: The linear Langevin system: RMSEs of 10^4 forecasting ensembles for different ensemble sizes $N_{ens} = 1$ (dash-dot line) and $N_{ens} = 20$ (solid line), produced by the true system (red cross marker), the system with estimated parameters (black triangle marker), and the ARMA model (blue circle marker). The ARMA model reproduces the RMSEs of the true system for all three step-sizes, while the estimated system has large RMSEs.

$$\mathrm{RMSE}(kh) := \left(\frac{1}{N_0} \sum_{i=1}^{N_0} \left| \bar{X}_k^i - x_{(n_i+k)h} \right|^2 \right)^{1=2} .$$

The RMSE measures the accuracy of the mean ensemble prediction; RMSE = 0 corresponds to a perfect prediction, and small RMSEs are desired.

The computed RMSEs for $N_0 = 10^4$ ensembles are shown in Figure 2, where we tested two ensemble sizes: $N_{ens} = 1, 20$. We observe that a larger ensemble size leads to smaller RMSEs for all the three systems, i.e. the ARMA$(2,1)$, the estimated Langevin equation and the true Langevin equation. The ARMA$(2,1)$ model reproduces almost exactly the RMSEs of the true Langevin system for all three observation step-sizes, while the estimated Langevin system has large RMSEs due to the biases in estimators and numerical instability. The steady increase in RMSE, even for the true system, is entirely expected because the

Table 3: Mean and standard deviation of the estimators of the parameters (, ,) of the Kramers equation in the continuous-time approach, computed on 100 simulations.

Estimator	True value	$h = 1/32$	$h = 1/16$	$h = 1/8$
$\hat{\ }$	0.5	0.8726 (0.0063)	1.2049 (0.0057)	1.7003 (0.0088)
$\hat{\ }$	0.3162	0.3501 (0.0007)	0.3662 (0.0007)	0.4225 (0.0009)
$\hat{\ }$	1	0.9964 (0.0014)	1.0132 (0.0027)	1.1150 (0.0065)

forecasting ensemble is driven by independent realizations of the forcing, as one cannot infer the white noise driving the system that originally generated the data.

4.2 The Kramers oscillator

We consider the Kramers equation in the following form

$$\begin{aligned} dx_t &= y_t dt, \\ dy_t &= (- \ y_t - \ ^{-2}x_t^3 + x_t)dt + \ dB_t, \end{aligned} \tag{4.2}$$

for which there are two potential wells located at $x = \pm$.

In the continuous-time approach, the estimator $\hat{\ } = \left(\hat{\ }, \hat{\ }, \hat{\ }\right)$ is given by

$$\hat{\ } = \arg \min_{=(\ , \ , \ )} \left[\sum_{n=1}^{N-3} \frac{3}{2} \frac{[\bar{x}_{n+2} - \bar{x}_{n+1} + h(\ \bar{x}_n + \ ^{-2}x_n^3 - x_n)]^2}{h \ ^2} + (N - 3) \log \ ^2 \right]. \tag{4.3}$$

As for the linear Langevin system case, we begin with investigating the reliability of the estimators. A hundred simulated datasets are generated from the above Kramers oscillator with true parameters $= 0.5$, $= 1/\sqrt{10}$, $= 1$, and with initial $x_0 = y_0 = 1/2$ and integration time interval $[0, 10^4]$. The estimators of (, ,) are computed for each dataset. Empirical mean and standard deviation of the estimators are shown in Table 3. We observe that the biases in the estimators increase as h increases, in particular, the estimator of $\hat{\ }$ has a very large bias.

For the discrete-time approach, we have to select one of the four NARMA$(2, q)$ models, Model (M1)–(M4). We make the selection ~~in the practical setting where we have~~ using data only from a single long trajectory (e.g. from the time interval $[0, T]$ with $T = 2^{18} \approx 2 \times 10^5$, and we use the first half of the data to estimate the parameters. We first estimate the parameters for each NARMA model with $q = 0$ and $q = 1$, using the conditional likelihood method described in Section 3.1. Then we make a selection by the criteria proposed in Section 3.1. First, we test numerical stability by running the model for a large time for different realizations of the noise sequence. We find that for our model, using the values of h tested here, Model (M1) is often numerically unstable, so we do not compare it to the other schemes here. (In situations where the Euler scheme is more stable, e.g., for smaller values of h or for other models, we would expect it to be useful as the basis of a NARMA approximation.)

k: repetitive

[I AM LOST IN THE NEXT PARAGRAPH- DO YOU MEAN YOU USE THE REMAINING HALF OF THE OBSERVATIONS TO TEST THE DATA? Yes, we use the 2nd half of data to test prediction.]

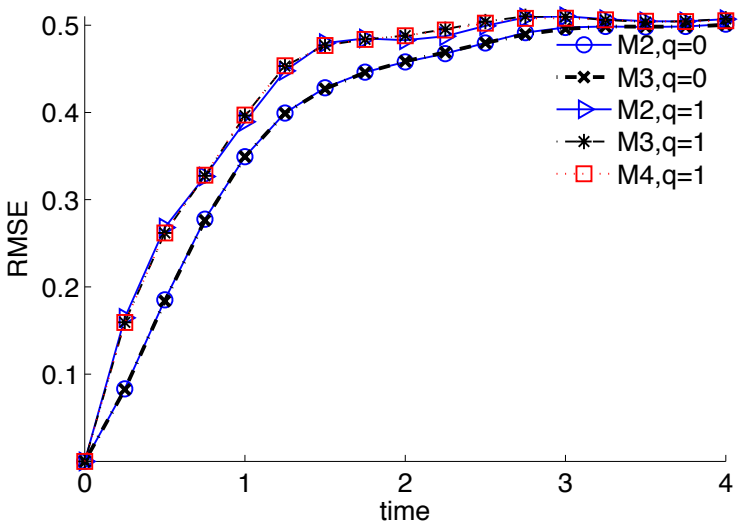


Figure 3: RMSEs of model (M2), (M3), (M4) with ensemble size $N_{ens} = 20$ in the case $h = 1/8$. Models with $q = 1$ have larger RMSEs than the models with $q = 0$. In the case $q = 0$, Model (M2) and (M3) have almost the same RMSEs.

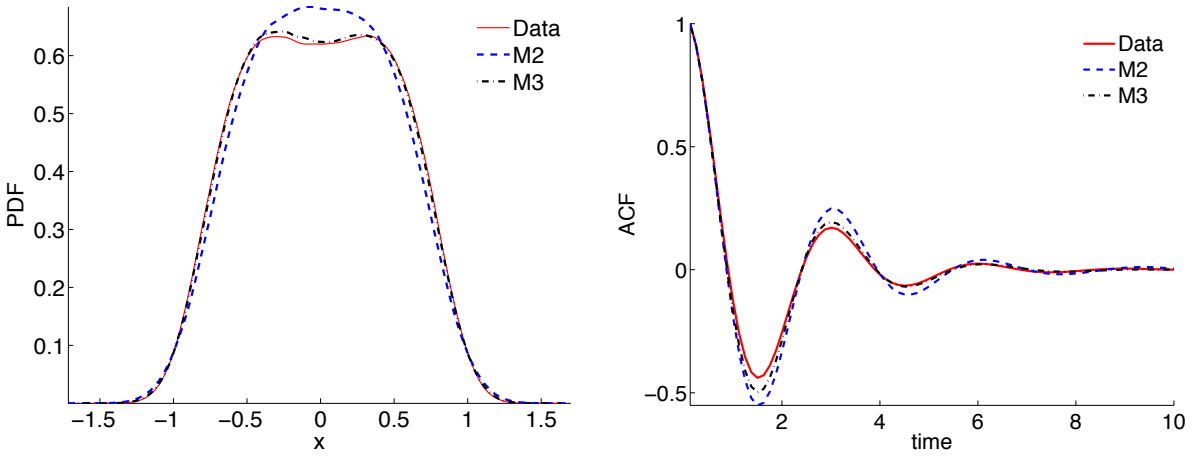
Next, ~~we test the predictions of the second half of the data using the remaining models~~ we test each of the models M2, M3, and M4 using the second half of the data. The RMSEs of models (M2), (M3) with $q = 0$ and $q = 1$ and Model (M4) with $q = 1$ are shown in Figure 3. In the case $q = 1$, the RMSEs for models (M2)-(M4) are very close, but they are larger than the RMSEs of models (M2) and (M3) in 4; this shows that model (M3) reproduces the ACFs and PDFs better than model (M2), hence model (M3) with $q = 0$ is selected.

The mean and standard deviation of estimated parameters of model (M3) with $q = 0$ and 100 simulations are shown in Table 4. Unlike the linear Langevin system case, we do not have a theoretical value for these parameters. However, note that when $h = 1/32$, $\hat{a}_1$ and $\hat{a}_2$ are close to $2 - h + {}^2h^2 = 1.9846$ and $-(1 - h + {}^2h^2) = -0.9846$ respectively, which are the coefficients in equation (3.7) from Itô-Taylor scheme. This indicates that when h is small, the NARMA model is close to the numerical scheme, because both the NARMA and the numerical scheme approximate the true system well. On the other hand, note that $\hat{}_W$ does not increase monotonically as h increases. This clearly distinguishes NARMA model from the numerical schemes.

Next, we test the performance of the NARMA model and the estimated Kramers system in reproducing long-term statistics and predicting short-term dynamics. The empirical PDFs and ACFs are shown in Figure 5. The NARMA models can reproduce the PDFs and ACFs equally well for three cases. The estimated Kramers system amplifies the depth of double wells in the PDFs, and it misses the oscillation of the ACFs.

k: Merged two paragraphs here.

Results for RMSEs for $N_0 = 10^4$ ensembles are shown in Figure 6, where we tested two ensemble sizes: $N_{ens} = 1, 20$. We observe that a larger ensemble size leads to smaller RMSEs for all the three systems, i.e. the NARMA model (M3) with $q = 0$, the estimated Kramers system and the true Kramers system. The NARMA model reproduces almost exactly the RMSEs of the true Kramers system for all three step-sizes, while the esti-

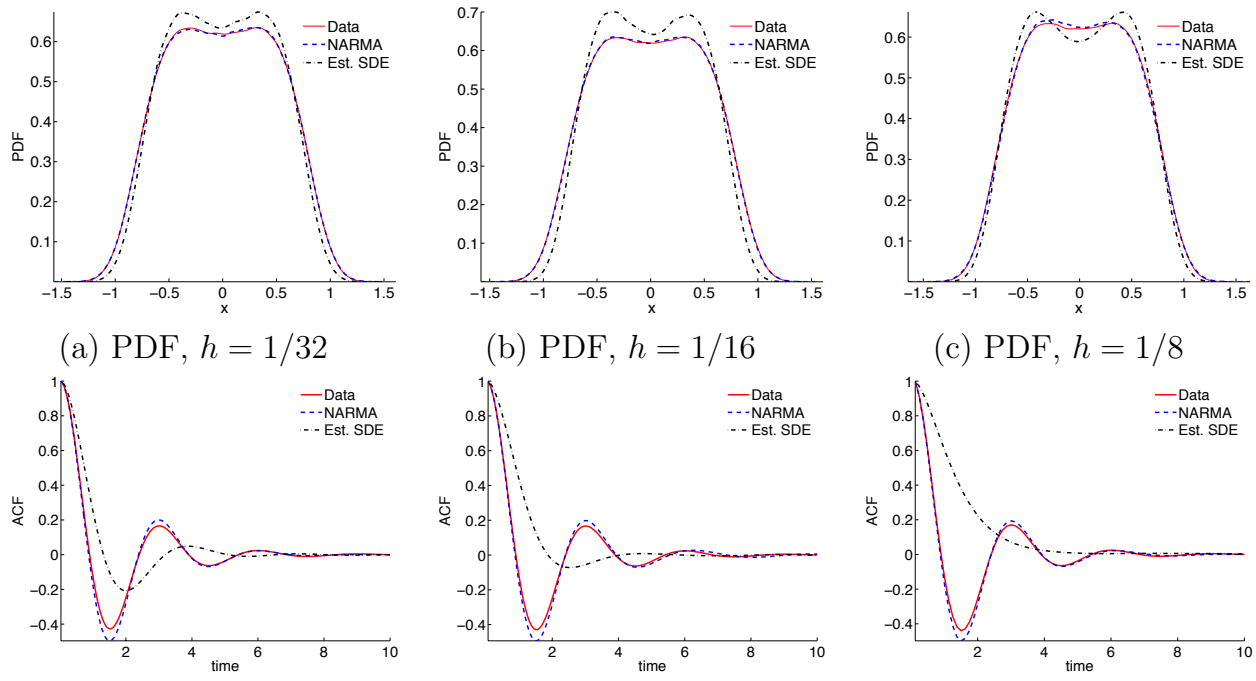


(a) PDF (b) ACF

Figure 4: Empirical PDFs and ACFs of the NARMA model (M2), (M3) and data in the case $h = 1/8$. Model (M3) reproduces the ACF and PDF better than model (M2).

Table 4: Mean and standard deviation of the estimators of the parameters $(a_1, a_2, b_1, b_2, b_3, \ , \ w)$ of the NARMA model (M3) with $q = 0$ in the discrete-time approach, computed from 100 simulations.

Estimator	$h = 1/32$	$h = 1/16$	$h = 1/8$
$\hat{a}_1$	1.9906 (0.0004)	1.9829 (0.0007)	1.9696 (0.0014)
$-\hat{a}_2$	0.9896(0.0004)	0.9792 (0.0007)	0.9562 (0.0014)
$-\hat{b}_1$	0.3388 (0.1572)	0.6927 (0.0785)	1.2988 (0.0389)
$\hat{b}_2$	0.0300 (0.1572)	0.0864 (0.0785)	0.1462 (0.0386)
$\hat{b}_3$	0.0307 (0.1569)	0.0887 (0.0777)	0.1655 (0.0372)
$-\hat{\ }$ $(\times 10^{-5})$	0.0377 (0.0000)	0.1478 (0.0000)	0.5469 (0.0001)
$\hat{\ }_w$	0.0045 (0.0000)	0.1119 (0.0001)	0.0012 (0.0000)



(a) PDF, $h = 1/32$ (b) PDF, $h = 1/16$ (c) PDF, $h = 1/8$

(d) ACF, $h = 1/32$ (e) ACF, $h = 1/16$ (f) ACF, $h = 1/8$

Figure 5: Empirical PDFs and ACFs of the NARMA model (M3) with q = 0 and the estimated Kramers system, in the cases h = 1=32, h = 1=16 and h = 1=8. These statistics are better reproduced by the NARMA models than the estimated Kramers systems.

mated Kramers system has increasing error as h increases, due to the increasing biases in estimators.

Finally, in Figure 7, we show some results using a much smaller observation spacing, h = 1=1024. Figure 7(a) shows the estimated parameters, for both the continuous-time and discrete-time models. (Here, the discrete-time model is M2.) Consistent with the theory in [ST12], our parameter estimates for the continuous time model are much closer to their true values for this smaller value of h. Figure 7(b) compares the RMSE of the continuous-time and discrete-time models on the same forecasting task as before. The continuous-time approach now performs much better, essentially as well as the true model. Even in this regime, however, the discrete-time approach remains competitive.

4.3 Discussion of structure design

[k: I do not understand the purpose of this section. I thought we asserted (in Section 3.1, after the definition of the estimator) that the estimator is consistent? What's the purpose of showing this data, if all we can conclude is that the estimators are likely consistent (which we already said) and we say conclusively that one model is better than another?]

In the above structure selection between model (M2) and (M3), we followed the criterion of selecting the one that fits the long-term statistics better. However, there is another practical criterion: the consistency of the estimators, i.e., whether the estimators converge to the true value as the number of samples tends to 1. Consistency can be tested by checking the oscillations of estimators as data length increases: if the oscillations are large,

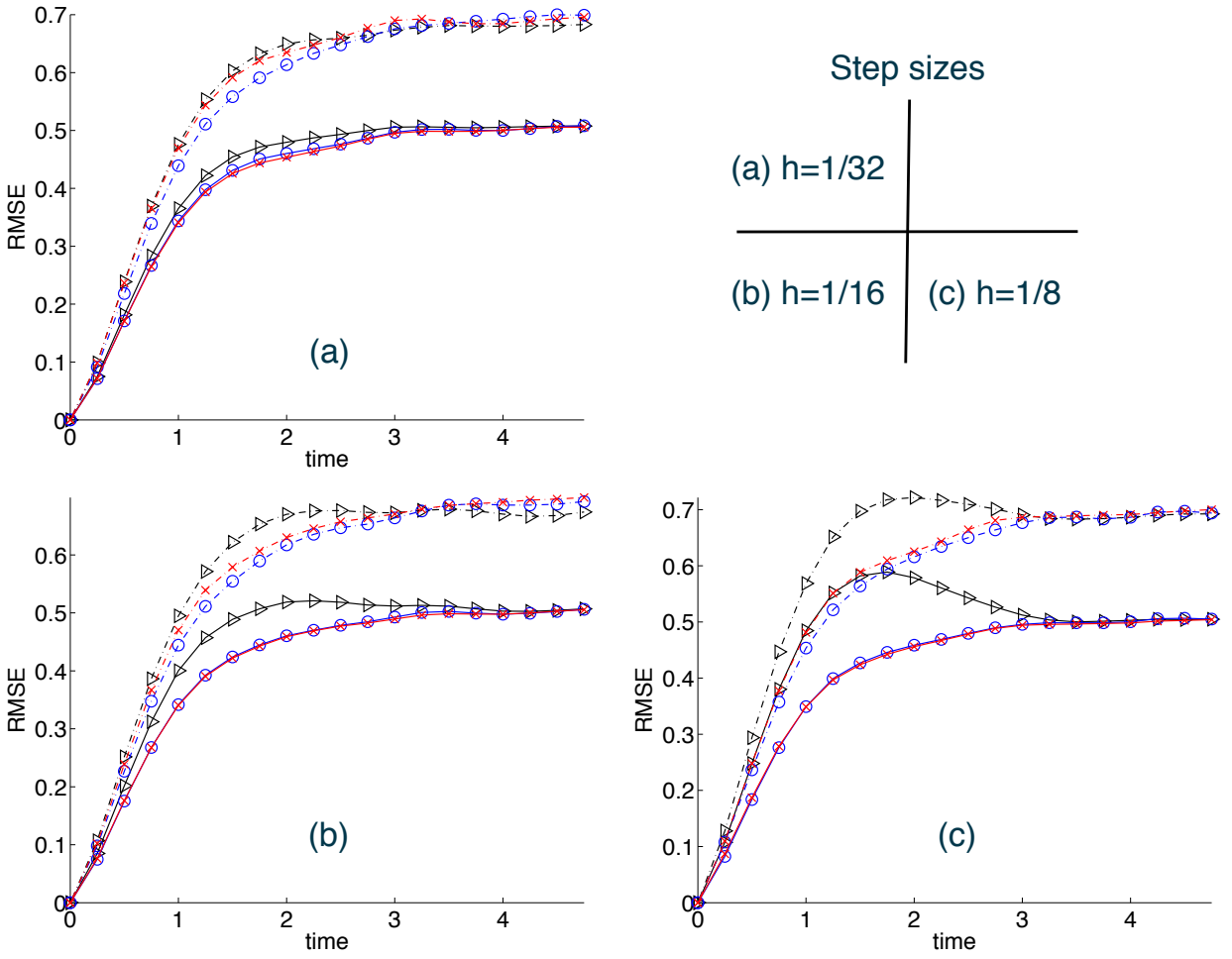


Figure 6: The Kramers system. (a) RMSEs of 10^4 forecasting ensembles for different ensemble sizes $N_{ens} = 1$ (dash-dot line) and $N_{ens} = 20$ (solid line), produced by the true Kramers system (red cross marker), the Kramers system with estimated parameters (black triangle marker), and the NARMA model (M3) with $q = 0$ (blue circle marker). The NARMA model has almost the same RMSEs as the true system for all three step-sizes, while the estimated system has larger RMSEs. (b) Estimated parameters for the continuous-time and discrete-time models.

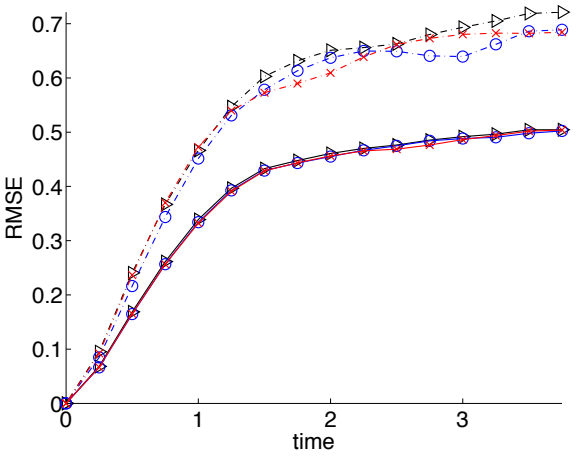
Table 5: Consistency test. [k: I removed the footnote: no need to repeat what's in the text.] Values of the estimators in the NARMA models (M2) and (M3) with $q = 0$. The data come from a long trajectory with observation step-size $h = 1/32$. Here $N = 2^{22}$. As the length of data increases, the estimators of Model (M2) have much smaller oscillation than the estimators of Model (M3).

Data length	Model (M2)		Model (M3)		
$(\times N)$	$-\hat{b}_1$	$-\hat{b}_2$	$-\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$
1=8	0.3090	0.3032	0.3622	0.0532	0.0563
1=4	0.3082	0.3049	0.3290	0.0208	0.0217
1=2	0.3088	0.3083	0.3956	0.0868	0.0845
1	0.3087	0.3054	0.3778	0.0691	0.0697

continuous time parameters		
$\hat{\ }$	$-\hat{\ }$	$\hat{\ }$
0.5163	0.3435	1.0006

DT parameters		
$\hat{a}_1$	$-\hat{a}_2$	$-\hat{b}_1$
1.9997	0.9997	0.0097
$-\hat{b}_2$	$-\hat{\ }(\times 10^{-8})$	$\hat{w}(\times 10^{-10})$
0.0169	2.0388	6.2165

(a) Estimated parameter values



(b) $h = 1/1024$

Figure 7: The Kramers system: RMSEs of 10^3 forecasting ensembles for different ensemble sizes $N_{ens} = 1$ (dash-dot line) and $N_{ens} = 20$ (solid line), produced by the true Kramers system (red cross marker), the Kramers system with estimated parameters (black triangle marker), and the NARMA model (M2) with $q = 0$ (blue circle marker). Since $h = 1/1024$ is relatively small, the NARMA model and the estimated system have almost the same RMSEs as the true system. Here the data is generated by the Itô-Taylor solver with step size $dt = 2^{-15} \approx 3 \times 10^{-5}$, and data length is $N = 2^{22} \approx 4 \times 10^6$.

the estimators are likely not to be consistent. Table 5 shows the estimators of the coefficients of the nonlinear terms in model (M2) and (M3), for different lengths of data. The estimators $\hat{b}_1, \hat{b}_2$ and $\hat{b}_3$ of model (M3) are unlikely to be consistent, since they vary a lot for long data sets. On the contrary, the estimators $\hat{b}_1$ and $\hat{b}_2$ of model (M2) have much smaller oscillations, and hence they are likely to be consistent.

These ~~results of~~ consistency tests agree with the statistics of the estimators in many simulations in Table 4 and Table 6. Table 4 shows that the standard deviations of the estimators $\hat{b}_1, \hat{b}_2$ and $\hat{b}_3$ are reduced by half as h doubles, which is the opposite of what is supposed to happen for an accurate model. On the contrary, Table 6 shows that the standard deviations of the parameters of model (M2) increase as h doubles, as is supposed to happen for an accurate model.

In short, model (M3) reproduces better long-term statistics than model (M2), but the estimators of model (M2) are statistically better (e.g. in consistency) than the estimators of model (M3). However, the two have almost the same prediction skill as shown in Figure 3, and both are much better than the continuous-time approach. It is unclear which model approximates the true process better, and it is likely that neither of them is optimal. Also, it is unclear which criterion is better for structure selection: fitting the long-term statistics or consistency of estimators. We leave these issues to be addressed in future work.

5 Conclusions

We have shown that for a prototypical hypoelliptic system, the discrete-time approach to data-driven prediction based on time-discrete partial observations generally has bet-

Table 6: Mean and standard deviation of the estimators of the parameters $(a_1, a_2, b_1, b_2, \ , \ _w)$ of the NARMA model (M2) with $q = 0$ in the discrete-time approach, computed on 100 simulations.

Estimator	$h = 1/32$	$h = 1/16$	$h = 1/8$
$\hat{a}_1$	1.9905 (0.0003)	1.9820 (0.0007)	1.9567 (0.0013)
$-\hat{a}_2$	0.9896 (0.0003)	0.9788 (0.0007)	0.9508 (0.0014)
$-\hat{b}_1$	0.3088 (0.0021)	0.6058 (0.0040)	1.1362 (0.0079)
$-\hat{b}_2$	0.3067 (0.0134)	0.5847 (0.0139)	0.9884 (0.0144)
$-\hat{} \ (\times 10^{-5})$	0.0340 (0.0000)	0.1193 (0.0000)	0.2620 (0.0001)
$\hat{}_w$	0.0045 (0.0000)	0.1119 (0.0001)	0.0012 (0.0000)

ter prediction skills than the continuous-time approach, especially when the time interval between observations is relatively large. We have also shown that the structure of the discrete-time model can be effectively derived from numerical schemes for solving SDEs. Since hypoelliptic systems are often used for the stochastic parametrization of data generated by high-dimensional dynamical systems, our findings suggest that a discrete-time approach may be both more efficient and more accurate than a continuous-time model, even when a parametric family containing the exact model is known.

Other advantages of the discrete approach are that its formulation does not produce a discretization error, so that it can tolerate large gaps between observation times, and does not require an additional discretization before it can be used, unlike a SDE.

In this paper we limited ourselves to a hypoelliptic system (1.1) with additive noise. Our discrete-time approach extends to multiplicative noise cases and general stochastic or deterministic ergodic dynamics systems. To deal with such systems, one needs to use a model of the following form

$$X_n = \quad (X_{n-p:n-1}, \ _{n-q:n-1}) + \quad (X_{n-p:n-1}, \ _{n-q:n-1}) \ _n,$$

where and are functions of $(X_{n-p:n-1}, \ _{n-q:n-1})$. The main change is in the structure derivation for the functions and from numerical schemes.

[k: This needs to be expanded a bit more, but maybe later.]

A Solutions to the linear Langevin equation

Denoting

$$X_t = \begin{pmatrix} x_t \\ y_t \end{pmatrix}, A = \begin{pmatrix} 0 & 1 \\ - & - \end{pmatrix}, e = \begin{pmatrix} 0 \\ \end{pmatrix},$$

we can write equation (3.2) as

$$dX_t = AX_t dt + e dB_t.$$

Its solution is

$$X_t = e^{At}X_0 + \int_0^{Z_t} e^{A(t-u)} e dB_u.$$

The solution at discrete times can be written as

$$x_{(n+1)h} = a_{11}x_{nh} + a_{12}y_{nh} + W_{n+1,1},$$
$$y_{(n+1)h} = a_{21}x_{nh} + a_{22}y_{nh} + W_{n+1,2},$$

where $a_{ij} = \left(e^{Ah}\right)_{ij}$ for $i, j = 1, 2$, and

$$W_{n+1,i} = \int_0^h a_{i2}(u)dB(nh + u) \tag{A.1}$$

with $a_{i2}(u) = \left(e^{A(h-u)}\right)_{i2}$ for $i = 1, 2$. Note that $a_{12} \neq 0$, then from the first equation we get $y_{nh} = \left(x_{(n+1)h} - a_{11}x_{nh} - V_{n+1,1}\right)/a_{12}$. Substituting it into the second equation we obtain

$$x_{(n+2)h} = (a_{11} + a_{22})\, x_{(n+1)h} + (a_{12}a_{21} - a_{11}a_{22})\, x_{nh}$$
$$- a_{22}W_{n+1,1} + a_{12}W_{n+1,2} + W_{n+2,1}.$$

Combining with the fact that $a_{11} + a_{22} = \mathrm{trace}(e^{Ah})$ and $a_{12}a_{21} - a_{11}a_{22} = -\det e^{Ah} = -e^{-\ h}$, we have

$$x_{(n+2)h} = \mathrm{trace}(e^{Ah})x_{(n+1)h} - e^{-\ h}x_{nh} - a_{22}W_{n+1,1} + W_{n+2,1} + a_{12}W_{n+1,2}. \tag{A.2}$$

Clearly, the process $\{x_{nh}\}$ is a centered Gaussian process, and its distribution is determined by its autocovariance function. Conditional on X_0, the distribution of X_t is $N(e^{At}X_0, \Sigma(t))$, where $\Sigma(t) := \int_0^t e^{Au}ee^{\mathsf{T}}e^{A^{\mathsf{T}}u}du$. Since $\ , \ > 0$, the real parts of the eigenvalues of the A, denoted by $\ _1$ and $\ _2$, are negative. The stationary distribution is $N(0, \Sigma(\ 1))$, where $\Sigma(\ 1) = \lim_{t\to\ 1} \Sigma(t)$. If X_0 has distribution $N(0, \Sigma(\ 1))$, then the process (X_t) is stationary, and so is the observed process $\{x_{nh}\}$. The following lemma computes the autocorrelation function of the stationary process $\{x_{nh}\}$.

Lemma A.1 *Assume that the system* (3.2) *is stationary. Denote by* $\{\ _j\}_{j=1}^{1}$ *the autocovariance function of the stationary process* $\{x_{nh}\}$*, i.e.* $\ _j := E[x_{kh}x_{(k+j)h}]$ *for* $j > 0$*. Then* $\ _0 = \frac{\ ^2}{2\ }$*, and* $\ _j$ *can be represented as*

$$\ _j = \ _0 \times \begin{cases} \frac{1}{\ _1 - \ _2}(\ _1 e^{\ _2 jh} - \ _2 e^{\ _1 jh}), & \text{if } \ ^2 - 4\ \neq 0; \\ e^{\ _0 jh}(1 - \ _0 jh), & \text{if } \ ^2 - 4\ = 0 \end{cases}$$

for all $j > 0$*, where* $\ _1$ *and* $\ _2$ *are the di erent solutions to* $\ ^2 + \ + \ = 0$ *when* $\ ^2 - 4\ \neq 0$*, and* $\ _0 = -\ /2$*.*

Proof. Let $\Gamma(j) := E[X_{kh}X_{(k+j)h}^{\mathsf{T}}] = \Sigma(\ 1)e^{A^{\mathsf{T}}jh}$ for $j > 0$. Note that $\ _j = \Gamma_{11}(j)$, i.e., $\ _j$ is the first element of the matrix $\Gamma(j)$. Then it follows that

$$\ _0 = \Sigma_{11}(\ 1), \quad \ _j = \left(\Sigma(\ 1)e^{A^{\mathsf{T}}jh}\right)_{11}.$$

If $\ ^2 - 4\ \neq 0$, then A has two different eigenvalues $\ _1$ and $\ _2$, and it can be written as

$$A = Q\Lambda Q^{-1} \text{ with } Q = \begin{pmatrix} 1 & 1 \\ \ _1 & \ _2 \end{pmatrix}, \Lambda = \begin{pmatrix} \ _1 & 0 \\ 0 & \ _2 \end{pmatrix}.$$

The covariance matrix $\Sigma(1)$ can be computed as

$$\Sigma(1) = \lim_{t\to 1} \int_0^t Q e^{\Lambda u} Q^{-1} e e^T Q^{-T} e^{\Lambda^T u} Q^T du = \sigma^2 \begin{pmatrix} \frac{1}{2ab} & 0 \\ 0 & -\frac{1}{2b} \end{pmatrix}. \tag{A.3}$$

This gives $\gamma_0 = \Sigma_{11}(1) = \frac{\sigma^2}{2}$ and for $j > 0$,

$$\gamma_j = \Sigma_{11}(1)\left(e^{A^T jh}\right)_{11} = \frac{1}{\lambda_1 - \lambda_2}(\lambda_1 e^{\lambda_2 jh} - \lambda_2 e^{\lambda_1 jh})\ (0).$$

In the case $\beta^2 - 4\alpha = 0$, A has a single eigenvalue $\lambda_0 = -\frac{\beta}{2}$, and it can be transformed to a Jordan block

$$A = Q\Lambda Q^{-1} \text{ with } Q = \begin{pmatrix} 1 & 0 \\ \lambda_0 & 1 \end{pmatrix}, \Lambda = \begin{pmatrix} \lambda_0 & 1 \\ 0 & \lambda_0 \end{pmatrix}.$$

This leads to the same $\Sigma(1)$ as in (A.3). Similarly, we have $\gamma_0 = \frac{\sigma^2}{2}$ and

$$\gamma_j = \Sigma_{11}(1)\left(e^{A^T jh}\right)_{11} = e^{\lambda_0 jh}(1 - \lambda_0 jh)\ \gamma_0.$$

■

B ARMA processes

We review the definition and computation of autocovariance function of ARMA processes in this subsection. For more details, we refer to [BD91, Section 3.3].

Definition B.1 *The process* $\{X_n, n \in \mathbb{Z}\}$ *is said to be an ARMA*(p,q) *process if it is stationary process satisfying*

$$X_n - \alpha_1 X_{n-1} - \cdots - \alpha_p X_{n-p} = W_n + \beta_1 W_{n-1} + \cdots + \beta_q W_{n-q}, \tag{B.1}$$

for every n, *where* $\{W_n\}$ *are i.i.d* $N(0, \sigma_W^2)$, *and if the polynomials* $\alpha(z) := 1 - \alpha_1 z - \cdots - \alpha_p z^p$ *and* $\beta(z) := 1 - \beta_1 z - \cdots - \beta_q z^q$ *have no common factors. If* $\{X_n - \mu\}$ *is an ARMA*(p,q) *process, then* $\{X_n\}$ *is said to be an ARMA*(p,q) *process with mean* μ. *The process is causal if* $\alpha(z) \neq 0$ *for all* $|z| \leq 1$. *The process is invertible if* $\beta(z) \neq 0$ *for all* $|z| \leq 1$.

The autocovariance function $\{\gamma(k)\}_{k=1}^{\infty}$ of an ARMA(p,q) can be computed from the following difference equations, which are obtained by multiplying each side of (B.1) by X_{n-k} and taking expectations,

$$\gamma(k) - \alpha_1 \gamma(k-1) - \cdots - \alpha_p \gamma(k-p) = \sigma_W^2 \sum_{k \leq j \leq q} \beta_j \psi_{j-k}, \quad 0 \leq k < \max\{p, q+1\}, \tag{B.2}$$

$$\gamma(k) - \alpha_1 \gamma(k-1) - \cdots - \alpha_p \gamma(k-p) = 0, \quad k > \max\{p, q+1\}, \tag{B.3}$$

where ψ_j in (B.2) is computed as follows (letting $\beta_0 := 1$ and $\beta_j = 0$ if $j > q$)

$$\gamma_j = \begin{cases} j + \sum_{0<k\le j} \alpha_k \gamma_{j-k}, & \text{for } j < \max\{p, q+1\}; \\ \sum_{0<k\le p} \alpha_k \gamma_{j-k}, & \text{for } j > \max\{p, q+1\}. \end{cases}$$

Denote $(\xi_i, i = 1, \dots, k)$ the distinct zeros of $\alpha(z) := 1 - \alpha_1 z - \cdots - \alpha_p z^p$, and let r_i be the multiplicity of ξ_i (hence $\sum_{i=1}^{k} r_i = p$). The general solution of the difference equation (B.3) is

$$\gamma(n) = \sum_{i=1}^{k} \sum_{j=0}^{r_i - 1} \beta_{ij} n^j \xi_i^{-n}, \text{ for } n > \max\{p, q+1\} - p, \tag{B.4}$$

where the p constants β_{ij} (and hence the values of $\gamma(j)$ for $0 \le j < \max\{p, q+1\} - p$) are determined from (B.2).

Example B.2 (ARMA$(2,0)$) . *For an ARMA(2,0) process* $X_n - \alpha_1 X_{n-1} - \alpha_2 X_{n-2} = W_n$, *its autocovariance function is*

$$\gamma(n) = \begin{cases} \beta_1 \xi_1^{-n} + \beta_2 \xi_2^{-n}, & \text{if } \alpha_1^2 + 4\alpha_2 \neq 0; \\ (\beta_1 + \beta_2 n)\xi_1^{-n}, & \text{if } \alpha_1^2 + 4\alpha_2 = 0 \end{cases}$$

for $n > 0$, *where* ξ_1, ξ_2 *or* ξ *are the zeros of* $\alpha(z) = 1 - \alpha_1 z - \alpha_2 z^2$. *The constants* β_1 *and* β_2 *are computed from the equations*

$$\begin{aligned} \gamma(0) - \alpha_1 \gamma(1) - \alpha_2 \gamma(2) &= \sigma_W^2, \\ \gamma(1) - \alpha_1 \gamma(0) - \alpha_2 \gamma(1) &= 0. \end{aligned}$$

Example B.3 (ARMA$(2,1)$) . *For an ARMA(2,1) process* $X_n - \alpha_1 X_{n-1} - \alpha_2 X_{n-2} = W_n + \theta_1 W_{n-1}$, *we have* $\psi_0 = 1$, $\psi_1 = \theta_1$. *Its autocovariance function is of the form as in* (B.2), *where the constants* β_1 *and* β_2 *are computed from the equations*

$$\begin{aligned} \gamma(0) - \alpha_1 \gamma(1) - \alpha_2 \gamma(2) &= \sigma_W^2(1 + \theta_1^2 + \theta_1 \psi_1), \\ \gamma(1) - \alpha_1 \gamma(0) - \alpha_2 \gamma(1) &= \sigma_W^2 \theta_1. \end{aligned}$$

C Numerical schemes for hypoelliptic SDEs with additive noise

Here we briefly review the two numerical schemes, the Euler-Maruyama scheme and the Itô-Taylor scheme of strong order 2.0, for hypoelliptic systems with additive noise

$$\begin{aligned} dx &= y\,dt, \\ dy &= a(x, y)dt + \sigma\, dB_t, \end{aligned}$$

where $a : R^2 \to R$ satisfies suitable conditions so that the system is ergodic.

In the following, the step size of all schemes are h, and $W_n = \sqrt{h}\,\xi_n$, $Z_n = h^{3/2}\left(\xi_n + \frac{1}{\sqrt{3}}\eta_n\right)$, where $\{\xi_n\}$ and $\{\eta_n\}$ are two i.i.d sequences of $N(0,1)$ random variables.

Euler-Maruyama (EM):

$$\begin{aligned} x_{n+1} &= x_n + y_n h, \\ y_{n+1} &= y_n + h a(x_n, y_n) + W_{n+1}. \end{aligned} \tag{C.1}$$

Itô-Taylor scheme of strong order 2.0 (IT2):

$$
\begin{aligned}
x_{n+1} &= x_n + h y_n + h^2 a(x_n, y_n) + Z_{n+1}, \\
y_{n+1} &= y_n + h a(x_n, y_n) + W_{n+1} + h^2 \left[a_x(x_n, y_n) y_n + \left(a a_y + \frac{1}{2} \sigma^2 a_{yy} \right)(x_n, y_n) \right] \\
&\quad + a_y(x_n, y_n) Z_{n+1} + a_{yy}(x_n, y_n) \sigma \frac{h}{6}(W_{n+1}^2 - h).
\end{aligned} \tag{C.2}
$$

The Itô-Taylor scheme of order 2.0 can be derived as follows (see e.g. Kloeden and Platen [Hu96, KP99]). The differential equation can be rewritten in the integral form:

$$
\begin{aligned}
x_t &= x_{t_0} + \int_{t_0}^{t} y_s ds, \\
y_t &= y_{t_0} + \int_{t_0}^{t} a(x_s, y_s) ds + \sigma (B_t - B_{t_0}).
\end{aligned}
$$

We start from Itô-Taylor expansion of x :

$$
\begin{aligned}
x_{t_{n+1}} &= x_{t_n} + h y_{t_n} + \int_{t_n}^{t_{n+1}} \int_{t_n}^{t} a(x_s, y_s) ds dt + I_{10}^{n+1} \\
&= x_{t_n} + h y_{t_n} + h^2 a(x_{t_n}, y_{t_n}) + \sigma I_{10}^{n+1} + O(h^{5/2}),
\end{aligned}
$$

where $I_{10}^{n+1} := \int_{t_n}^{t_{n+1}} (B_t - B_{t_n})\, dt$. To get higher order scheme for y, we apply Itô's chain rule to $a(x_t, y_t)$:

$$
a(x_t, y_t) = a(x_s, y_s) + \int_s^t \left[a_x(x_r, y_r) y_r + \left(a a_y + \frac{1}{2} \sigma^2 a_{yy} \right)(x_r, y_r) \right] dr + \int_s^t \sigma a_y(x_r, y_r)\, dB_r.
$$

This leads to Itô-Taylor expansion for y (up to the order 2.0):

$$
\begin{aligned}
y_{t_{n+1}} &= y_{t_n} + \int_{t_n}^{t_{n+1}} a(x_s, y_s) ds + \sigma (B_{t_{n+1}} - B_{t_n}) \\
&= y_{t_n} + h a(x_{t_n}, y_{t_n}) + \sigma (B_{t_{n+1}} - B_{t_n}) \\
&\quad + h^2 \left[a_x(x_{t_n}, y_{t_n}) y_{t_n} + \left(a a_y + \frac{1}{2} \sigma^2 a_{yy} \right)(x_{t_n}, y_{t_n}) \right] \\
&\quad + a_y(x_{t_n}, y_{t_n}) \sigma I_{10}^{n+1} + a_{yy}(x_{t_n}, y_{t_n}) \sigma^2 I_{110}^{n+1} + O\left(h^{5/2}\right).
\end{aligned}
$$

where $I_{110}^{n+1} = \int_{t_n}^{t_{n+1}} \int_{t_n}^{t} (B_s - B_{t_n})\, dB_s dt$. Representing $\sigma (B_{t_{n+1}} - B_{t_n})$, σI_{10}^{n+1} and σI_{110}^{n+1} by W_{n+1}, Z_{n+1} and $\frac{h}{6}(W_{n+1}^2 - h)$ respectively, we obtain scheme (C.2).

References

[AI00] L. Arnold and P. Imkellerf. The Kramers oscillator revisited. In J. Freund and T. Pöschel, editors, *Stochastic Processes in Physics, Chemistry, and Biology*, volume 557 of *Lecture Notes in Physics*, page 280. Springer, Berlin, 2000.

[AM11] D. F. Anderson and J. C. Mattingly. A weak trapezoidal method for a class of stochastic differential equations. *Commun. Math. Sci.*, 9(1), 2011.

[And70] E. B. Andersen. Asymptotic properties of conditional maximum-likelihood estimators. *J. R. Stat. Soc. Series B*, pages 283–301, 1970.

[BD91] P. Brockwell and R. Davis. *Time series: theory and methods*. Springer, New York, 2nd edition, 1991.

[BDY07] P.J. Brockwell, R. Davis, and Y. Yang. Continuous-time Gaussian autoregression. *Statistica Sinica*, 17(1):63, 2007.

[Bro01] P.J. Brockwell. Continuous-time ARMA processes. *Handbook of Statistics*, 19:249–276, 2001.

[Bro14] P.J. Brockwell. Recent results in the theory and applications of CARMA processes. *Ann. Inst. Stat. Math.*, 66(4):647–685, 2014.

[CL15] A.J. Chorin and F. Lu. Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics. *Proc. Natl. Acad. Sci. USA*, 112(32):9804–9809, 2015.

[DS04] S. Ditlevsen and M. Sørensen. Inference for observations of integrated diffusion processes. *Scand. J. Statist.*, 31(3):417–429, 2004.

[FS01] D. Frenkel and B. Smit. *Understanding molecular simulation: from algorithms to applications*, volume 1. Academic press, 2001.

[GCF15] G.A. Gottwald, D. Crommelin, and C. Franzke. Stochastic climate theory. In *Nonlinear and Stochastic Climate Dynamics*. Cambridge University Press, 2015.

[Glo06] A. Gloter. Parameter estimation for a discretely observed integrated diffusion process. *Scand. J. Statist.*, 33(1):83–104, 2006.

[Ham94] J.D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, NJ, 1994.

[Hu96] Y. Hu. Strong and weak order of time discretization schemes of stochastic differential equations. In *Séminaire de Probabilités XXX*, pages 218–227. Springer, 1996.

[Hum05] G. Hummer. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.*, 7(1):34, 2005.

[Jen14] Anders Christian Jensen. *Statistical Inference for Partially Observed Diffusion Processes*. PhD thesis, University of Copenhagen, Faculty of Science, Department of Mathematical Sciences, 2014.

[Jon81] R. H. Jones. Jones fitting a continuous time autoregressive to discrete data. In D. F. Findley, editor, *Applied Time Series Analysis II*, pages 651–682. Academic Press, New York, 1981.

[KP99] P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Di erential Equations*. Springer, Berlin, 3rd edition, 1999.

[Kra40] H. A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

[LLC15] F. Lu, K. Lin, and A.J. Chorin. Data-based stochastic model reduction for the Kuramoto–Sivashinsky equation. *arXiv:1509.09279*, 2015.

[MH13] A.J. Majda and J. Harlim. Physics constrained nonlinear regression models for

time series. *Nonlinearity*, 26(1):201–217, 2013.

[MSH02] J.C. Mattingly, A.M. Stuart, and D.J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Process. Appl.*, 101:185–232, 2002.

[MST10] J.C. Mattingly, A.M. Stuart, and M.V. Tretyakov. Convergence of numerical time-averaging and stationary measures via Poisson equations. *SIAM J. Numer. Anal.*, 48(2):552–577, 2010.

[MT07] G.N. Milstein and M.V. Tretyakov. Computing ergodic limits for Langevin equations. *Physica D: Nonlinear Phenomena*, 229(1):81–95, 2007.

[Nua06] D. Nualart. *The Malliavin calculus and related topics*. Springer-Verlag, 2nd edition, 2006.

[Phi59] A.W. Phillips. The estimation of parameters in systems of stochastic differential equations. *Biometrika*, 46(1-2):67–76, 1959.

[PSW09] Y. Pokern, A.M. Stuart, and P. Wiberg. Parameter estimation for partially observed hypoelliptic diffusions. *J. Roy. Statis. Soc. B*, 71(1):49–73, 2009.

[Rao99] P. B.L.S. Rao. *Statistical Inference for Di usion Type Processes*. Oxford University Press, 1999.

[SGH93] L. Schimansky-Geier and H. Herzel. Positive lyapunov exponents in the Kramers oscillator. *J. Stat. Phys.*, 70(1-2):141–147, 1993.

[Sør04] H. Sørensen. Parametric inference for diffusion processes observed at discrete points in time: a survey. *International Statistical Review*, 72(3):337–354, 2004.

[Sør12] M. Sørensen. Estimating functions for diffusion-type processes. In M. Kessler, A. Lindner, and M. Sørensen, editors, *Statistical Methods for Stochastic Differential Equations*. Oxford University Press, London, 2012.

[ST12] A. Samson and M. Thieullen. A contrast estimator for completely or partially observed hypoelliptic diffusion. *Stochastic Process. Appl.*, 122(7):2521–2552, 2012.

[Tal02] D. Talay. Stochastic Hamiltonian systems: exponential convergence to the invariant measure, and discretization by the implicit Euler scheme. *Markov Process. Related Fields*, 8(2):163 – 198, 2002.