HAL open science

Building General Langevin Models from Discrete Datasets

Federica Ferretti, Victor Chardès, Thierry Mora, Aleksandra Walczak, Irene Giardina

▶ **To cite this version:**

Federica Ferretti, Victor Chardès, Thierry Mora, Aleksandra Walczak, Irene Giardina. Building General Langevin Models from Discrete Datasets. Physical Review X, 2020, 10 (3), 10.1103/PhysRevX.10.031018 . hal-02988463

HAL Id: hal-02988463

https://hal.science/hal-02988463v1

Submitted on 5 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Building General Langevin Models from Discrete Datasets

Federica Ferretti,[1,2] Victor Chardès,[3] Thierry Mora,[3] Aleksandra M. Walczak,[3] and Irene Giardina[1,2,4]

[1]*Dipartimento di Fisica, Università Sapienza, 00185 Rome, Italy*
[2]*Istituto Sistemi Complessi, Consiglio Nazionale delle Ricerche, UOS Sapienza, 00185 Rome, Italy*
[3]*Laboratoire de Physique de l'École Normale supérieure (PSL University), CNRS, Sorbonne Université, Université de Paris, 75005 Paris, France*
[4]*INFN, Unità di Roma 1, 00185 Rome, Italy*

(Received 23 December 2019; revised 18 April 2020; accepted 22 May 2020; published 23 July 2020)

Many living and complex systems exhibit second-order emergent dynamics. Limited experimental access to the configurational degrees of freedom results in data that appear to be generated by a non-Markovian process. This limitation poses a challenge in the quantitative reconstruction of the model from experimental data, even in the simple case of equilibrium Langevin dynamics of Hamiltonian systems. We develop a novel Bayesian inference approach to learn the parameters of such stochastic effective models from discrete finite-length trajectories. We first discuss the failure of naive inference approaches based on the estimation of derivatives through finite differences, regardless of the time resolution and the length of the sampled trajectories. We then derive, adopting higher-order discretization schemes, maximum-likelihood estimators for the model parameters that provide excellent results even with moderately long trajectories. We apply our method to second-order models of collective motion and show that our results also hold in the presence of interactions.

DOI: 10.1103/PhysRevX.10.031018

Subject Areas: Computational Physics, Statistical Physics

I. INTRODUCTION

Recent experimental findings on a variety of living systems, from cell migration [1], bacterial propulsion [2], and worm dynamics [3] to the larger scale of animal groups on the move [4–7], indicate that the observed behavior cannot be explained with a first-order dynamical process but requires a higher-order description. For bird flocks and insect swarms, the case which interests us most, data show that propagating directional information during collective turns in flocks requires rotational inertia, i.e., a reversible dynamical term, to account for the measured dispersion law [6]. The shape of the velocity-velocity correlation function in swarms, which flattens at short times, also points to a second-order dynamics for these systems, as suggested by the value of the dynamical critical exponent [7]. Overall, data indicate that considering second-order dynamics is required to explain how animal groups behave on their natural size and timescales—even though overdamping might theoretically occur for very large systems and on very large timescales.

The emergent dynamics of all the above systems share three fundamental ingredients: an effective inertia, dissipation, and a stochastic contribution. Disentangling such contributions is often crucial to understand the processes at stake, and reliable methods are required to extract that information from available data. The example of animal groups, which motivates the present work, is also helpful to discuss the theoretical objectives and experimental constraints of the inference procedure. Ideally, we would like to build the simplest continuous second-order model consistent with experimental findings. We seek a continuous-time model for several reasons: (i) It allows computations to be performed; (ii) it is a reasonable assumption for systems where microscopic update times are much smaller than observational scales (cognitive processes occur on tenths of milliseconds, whereas behavioral changes on scales of seconds); (iii) it circumvents the inherent arbitrariness of discrete-time modeling. Experimental data, on the other hand, come in the form of discrete timeseries, where the discretization interval is set by the time resolution of the experimental apparatus.

In the presence of stochasticity, the nature of the data poses two major problems. First of all, if the dynamics is of second order, all signals (including initial condition and noise) are propagated in time with a memory kernel, making the relation between the coarse-grained data that we observe and the underlying process far more complex

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

than in the first-order case. The memory kernel arises from the contraction of the dynamical description of the second-order stochastic process from the full phase space to a lower-dimensional subspace—usually that of measurable degrees of freedom [8–10]. For example, were we able to experimentally measure with the same accuracy a pair of conjugate variables, e.g., positions and velocities of moving individuals, we could seek a model for their joint evolution. But in common experiments that is not the case, as one typically measures one degree of freedom (e.g., positions) and must derive the other. To confront the data, we therefore need to work in a reduced space. Second, the goal of the inference procedure is to retrieve a continuous stochastic model from a collection of discrete sample paths occurring on finite observational timescales Δt. In the absence of an explicit solution for the stochastic process, the most reasonable thing to do is to transform the stochastic differential equation (SDE) into an approximated difference equation. Such discretization must be performed very carefully, since the resulting equation should correctly represent the underlying stochastic process both at the scales of the sampled data (at which inference works) and in the microscopic limit of vanishing increments.

These two problems are quite general and do not depend on the presence of interactions in the system but rather on the nature of the dynamics. Although the issue has been considered before, the literature is sparse and a satisfying Bayesian inference approach is still lacking. Previous attempts to provide systematic inference strategies for second-order dynamics consist of building converging estimators for the different terms of the model from proper combinations of measurable quantities [11–13] or in exploiting known relations between model parameters and accessible observables [14].

In a more general and refined way, the problem can be reformulated in terms of a dynamical inverse problem, and much work has been done in this field in the past years [15–17]. However, most analyses focus on first-order processes in time [18–25]. Second-order processes are considered within this framework in Refs. [26,27], yet the proposed method differs from a proper maximum-likelihood approach, due to the difficulty of deducing a pseudolikelihood function in the case of non-Markovian processes.

To the best of our knowledge, we present here the first maximum-likelihood inference approach for non-Markovian inertial processes. It differs from previous studies in its first principle derivation and the absence of a Markovian embedding. We derive explicit formulas for the parameter estimators and test our approach on synthetic data in a variety of models, including nonlinear forces, multiplicative noise, and many-body interactions. Results show that the method is accurate and robust, providing an important tool in the analysis and understanding of real systems. The paper is organized in the following way: In Sec. II, we formalize the problem and discuss in detail how to build an appropriate dynamical inference strategy for inertial systems with linear dissipation. We explain the interplay between the order of convergence of discretized SDEs obtained from Taylor-Itô expansions and the consistency of the corresponding max-likelihood parameter estimators. We show that to get accurate results the simplest Euler-like schemes, which work well with first-order dynamics, are insufficient, so that one needs to go to the next order of approximation. Theoretical predictions are compared with numerical data to consolidate our results. Section III introduces non-Bayesian inference schemes, while in Sec. IV we discuss the problem of eliminating the initial velocity. In Sec. V, we address the case of a strongly interacting system: The inference procedure is applied to synthetic data obeying the inertial spin model, a model of self-propelled particles that describes the phenomenology of natural flocks of birds [6]. The effect of experimental measurement noise is discussed in Sec. VI. Finally, in Sec. VII, we summarize all our results, discuss their conceptual relevance, and outline their potential for applications to real data.

II. MAXIMUM-LIKELIHOOD INFERENCE APPROACH FOR LANGEVIN DYNAMICS

A. Problem definition

Let us assume that the available experimental data are sequences of points $(x_0, x_1, \ldots, x_L)$ uniformly separated in time by Δt and that the underlying dynamics is described by the complete Langevin equation of the form

$$\ddot{x} = -\eta\dot{x} + f(x) + \sigma\xi, \tag{1}$$

where $f(x) = -V'(x)$ is a conservative force, $\sigma^2 = 2T\eta$, and ξ is a standard white noise: $\langle\xi\rangle = 0$, $\langle\xi(t)\xi(t')\rangle = \delta(t-t')$. Without a lack of generality, the inertial mass is set to 1. Since the noise is additive, it is unnecessary to distinguish between Itô and Stratonovich integration.

Let us call $\boldsymbol{\lambda}$ the irreducible set of parameters that enter in Eq. (1), namely, the effective damping coefficient η, the effective temperature T, and the parameters entering in the definition of the potential $V(x)$. The aim of dynamical statistical inference is to provide an estimate of their values. Following a Bayesian approach, the posterior distribution of parameters given the data reads

$$P(\boldsymbol{\lambda}|\{(x_0, \ldots, x_L)^\alpha\}) \propto P(\{(x_0, \ldots, x_L)^\alpha\}|\boldsymbol{\lambda})\rho(\boldsymbol{\lambda}), \tag{2}$$

where each Greek index labels a different experimental sample. By choosing a uniform prior $\rho(\boldsymbol{\lambda})$, the maximum of Eq. (2) corresponds to the maximum-likelihood estimator. The conceptual and technical difficulty of the whole inference problem is then only about finding a tractable expression for the dynamical likelihood.

The theory of stochastic processes provides us with an explicit but formal expression for the transition probability

$P[x(t)|x(0), \dot{x}(0)]$, involving, in general, integro-differential operators. A closed form solution for the stochastic process may be generally unknown or complicated [28], especially for many-body or off-equilibrium systems, but finely time-resolved data may be available. What we look for is then an (eventually approximated) expression for the probability of the discrete trajectory, for which a practical connection with the data can be established.

A first general strategy is the following:

(1) As a preliminary step, Eq. (1) can be conveniently rewritten as a set of two first-order equations:

$$\dot{x} = v, \qquad \dot{v} = -\eta v + f(x) + \sigma\xi. \tag{3}$$

(2) Since the dynamics is Markovian when parametrized by the vector variable $\mathbf{q} = (x, v)$, the probability of a discrete trajectory in this space, given the initial condition $\mathbf{q}_0 = (x_0, v_0)$, can be split into a product of propagators:

$$P(\mathbf{q}_L, \ldots, \mathbf{q}_1|\mathbf{q}_0) = \prod_{n=1}^{L} P(\mathbf{q}_n|\mathbf{q}_{n-1}). \tag{4}$$

(3) Following Ref. [30], one can exploit any update rule based on a Taylor-Itô expansion to approximate, within a certain order of accuracy, the propagator over a small time interval Δt:

$$P(\mathbf{q}_n|\mathbf{q}_{n-1}) = P_{(k)}(\mathbf{q}_n|\mathbf{q}_{n-1}) + o(\Delta t^k). \tag{5}$$

Equation (5) can be replaced into Eq. (4) to get an approximated expression for the probability density of the sequence of points in phase space:

$$P_{(k)}(\mathbf{q}_L, \ldots, \mathbf{q}_1|\mathbf{q}_0) = \prod_{n=1}^{L} P_{(k)}(\mathbf{q}_n|\mathbf{q}_{n-1}) + o(\Delta t^k). \tag{6}$$

(4) Marginalizing over the velocitylike degrees of freedom, one gets a probability distribution depending on the x's only. This projection operation on the subspace of x variables is where the original Markovian property of Eq. (4) is generally lost. A crucial remark, beyond the non-Markovian nature of the resulting dynamics, is that this procedure does not simply consist of removing the intermediate variables $v_1, \ldots, v_L$, but also of eliminating the initial condition v_0. This elimination is at the same time a further technical difficulty and a fundamental conceptual issue in the context of stochastic dynamics. We refer to Sec. IV for a broader discussion.

When this strategy is adopted, the first thing we need is then a discrete integration scheme for Eq. (1) or (3). Although the naive intuition is that any convergent—even if slowly—discretization scheme should work for small Δt, in fact, the order of approximation of the temporal discretization is able to affect the mathematical properties of the discrete path integral measure and, consequently, the correctness of estimators obtained through a maximum-likelihood inference procedure [26,30].

Alternatively, one can follow a second strategy, summarized as "*first marginalize, then discretize*," in contrast to the "*first discretize, then marginalize*" strategy discussed above. The starting point is here the generalized Langevin equation (GLE) corresponding to the desired process [Eq. (1)], which can be obtained adopting the Mori-Zwanzig formalism [8] (see the Appendix A):

$$\dot{x} = v_0 e^{-\eta t} + \int_0^t ds K(t-s) f(x(s)) + \zeta(t). \tag{7}$$

In this equation, $K(t) = e^{-\eta t}$, and the effective noise is given by $\zeta(t) = \int_0^t ds e^{-\eta(t-s)}\xi(s)$. This formalism shows that, when projected from the full phase space into the x space, the dynamics acquires a memory, described by a friction kernel $K(t)$ and color in the noise. We note that the relation $\langle\zeta(t)\zeta(t')\rangle \propto K(|t-t'|)$ holds asymptotically in the limit of an infinitely long trajectory, and it reduces to the second fluctuation dissipation theorem when $f(x)$ is linear. Discrete update equations can now be obtained by integrating Eq. (7) on Δt intervals and self-consistently removing v_0. We notice that, for arbitrary forces $f(x)$, the corresponding term cannot be exactly integrated, and it needs to be approximated at small Δt. The fact that the derivative of the measured coordinate—position x—enters parametrically through v_0 in the GLE stems from the second-order nature of the process. Its elimination, which is necessary to retrieve a stochastic difference equation where only the x variable appears, is connected to the problem anticipated in point 4 of the procedure outlined above.

The two strategies must be equivalent: The order of the discretization and marginalization operations should be exchangeable. In the following section, we show how the simplest inference schemes derived from Euler-like discretizations of Eq. (3) do not satisfy this requirement, whereas higher-order discretization schemes, strongly convergent as at least $O(\Delta t^{3/2})$, retrieve correct maximum-likelihood estimators.

B. Failure of naïve inference schemes

Discrete integration approaches for SDEs are well known in the literature in connection to numerical computation methods (see, e.g., Ref. [31]). Here, we summarize how the order of approximation of these discretization schemes interferes with the non-Markovian character of the observed dynamics. Because of this fact, standard claims about the convergence of the adopted integrators are not generally valid in cases when only a projection of the original Markovian process is observed. Rigorous results can be found in Ref. [26]. We are mainly interested in, from an application point of view, the bias that this fact

introduces in naïve inference approaches and possible correction strategies.

Let us start then with the simplest possible construction, i.e., the Euler-Maruyama scheme applied to Eq. (1) (in this case, corresponding to the Milstein scheme) [31]. The discrete update equations for the Markov process read

$$x_{n+1} - x_n = \Delta t v_n,$$
$$v_{n+1} - v_n = -\eta \Delta t v_n - \Delta t f(x_n) + \sigma \Delta t^{1/2} r_n, \quad (8)$$

with r_n independent identically distributed (IID) random variables of normal distribution $\mathcal{N}(0,1)$, for $n = 0, \ldots, L-1$. We recall that the first neglected terms in Eq. (8) are $O(\Delta t^{3/2})$. The scheme provides then a deterministic update for the x variables, which manifests itself through δ functions; a simple change of variables from r_n to v_{n+1} immediately completes the derivation of the discrete propagator in (x, v) space. Finally, in this case, one can explicitly marginalize over the velocity degrees of freedom and eliminate the initial condition v_0. Indeed, to this order of approximation, information on v_0 is fully equivalent to information on x_1. From this marginalization, a fully factorized probability distribution for the discrete sequence is obtained:

$$P_{(1)}(x_L, \ldots, x_2 | x_0, x_1) = \prod_{n=1}^{L-1} P_{(1)}(x_{n+1} | x_n, x_{n-1}), \quad (9)$$

where transition probabilities are defined as follows:

$$P_{(1)}(x_{n+1} | x_n, x_{n-1}) = \frac{1}{Z_n} e^{-S_n(x_{n+1}, x_n, x_{n-1})}, \quad (10)$$

with

$$Z_n = \sqrt{2\pi\sigma^2 \Delta t^3}; \quad (11)$$

$$S_n = \frac{1}{2\sigma^2 \Delta t^3} [x_{n+1} - 2x_n + x_{n-1} + \eta \Delta t (x_n - x_{n-1}) - \Delta t^2 f(x_n)]^2. \quad (12)$$

A factorization of $P(x_L, \ldots, x_2 | x_1, x_0)$ into a product of transition probabilities of this kind is possible because the random variables appearing in the x difference equation, obtained from Eq. (8) through variable elimination, are independent. This fact is a crucial but artificial feature occurring only at this level of approximation: More accurate discretization procedures produce an effective noise for the x variables which is correlated in time. As a matter of fact, when the description of a Brownian motion is contracted from the full phase space to position space, a colored noise emerges, which is incompatible with the independence of subsequent random variables at any Δt.

Nonetheless, we find it useful to compute the associated dynamical likelihood, as defined in Eq. (2), and develop the corresponding inference scheme. For the sake of clarity, we focus on the example of the harmonic oscillator, where $f(x) = -\omega_0^2 x$. Using Eqs. (10)–(12), an expression for the likelihood as a product of transition probabilities for a second-order master equation is recovered. This result corresponds to the discrete path probability one would obtain adopting a maximum caliber approach [15] when certain time-dependent observables are taken as fixed. For the one-dimensional harmonic oscillator, they are the equal-time correlations, one-time-step correlations, and two-time-step correlations of the process. Indeed, rearranging the sum of S_n's in Eq. (9), the reduced minus-log-likelihood can be written as

$$\frac{\mathcal{L}(\eta, T, \omega_0^2)}{L-1} = \frac{1}{2}\ln(2\pi\sigma^2\Delta t^3) + \frac{1}{2\sigma^2\Delta t^3}[C_s' + (2 - \eta\Delta t + \omega_0^2\Delta t^2)^2 C_s + (1-\eta\Delta t)^2 C_s'' + 2(1-\eta\Delta t)F_s - 2(2-\eta\Delta t + \omega_0^2\Delta t^2)G_s - 2(1-\eta\Delta t)(2-\eta\Delta t+\omega_0^2\Delta t^2)G_s'], \quad (13)$$

where we introduce the following notation for the experimental temporal correlation functions, evaluated at a time distance of 0, Δt, and $2\Delta t$:

$$C_s = \frac{1}{L-1}\sum_{n=1}^{L-1} x_n x_n; \qquad C_s' = \frac{1}{L-1}\sum_{n=1}^{L-1} x_{n+1} x_{n+1};$$
$$C_s'' = \frac{1}{L-1}\sum_{n=1}^{L-1} x_{n-1} x_{n-1}; \qquad G_s = \frac{1}{L-1}\sum_{n=1}^{L-1} x_n x_{n+1};$$
$$G_s' = \frac{1}{L-1}\sum_{n=1}^{L-1} x_n x_{n-1}; \qquad F_s = \frac{1}{L-1}\sum_{n=1}^{L-1} x_{n-1} x_{n+1}.$$

Minimization of the quantity in Eq. (13) with respect to η, T, and ω_0^2 yields the inference formulas for the parameters of the harmonic oscillator. We express here only the estimator of the damping coefficient η, while the remaining ones can be found in Appendix B 1:

$$\eta^* = \frac{1}{\Delta t}\frac{2C_s - G_s - G_s' - \frac{G_s'}{C_s''}(2G_s' - C_s'' - F_s)}{C_s + C_s'' - 2G_s' - \frac{(C_s'' - G_s')^2}{C_s''}}. \quad (14)$$

At this point, having an explicit inference method, it can be both numerically and analytically tested. We simulate discrete trajectories of the stochastic harmonic oscillator in several damping conditions using an exact integrator [32], with a numerical time step $\tau^{\mathrm{sim}} = 0.005$. We apply inference formulas to discrete datasets sampled from synthetic trajectories at time intervals $\Delta t \geq \tau^{\mathrm{sim}}$. This choice mimics

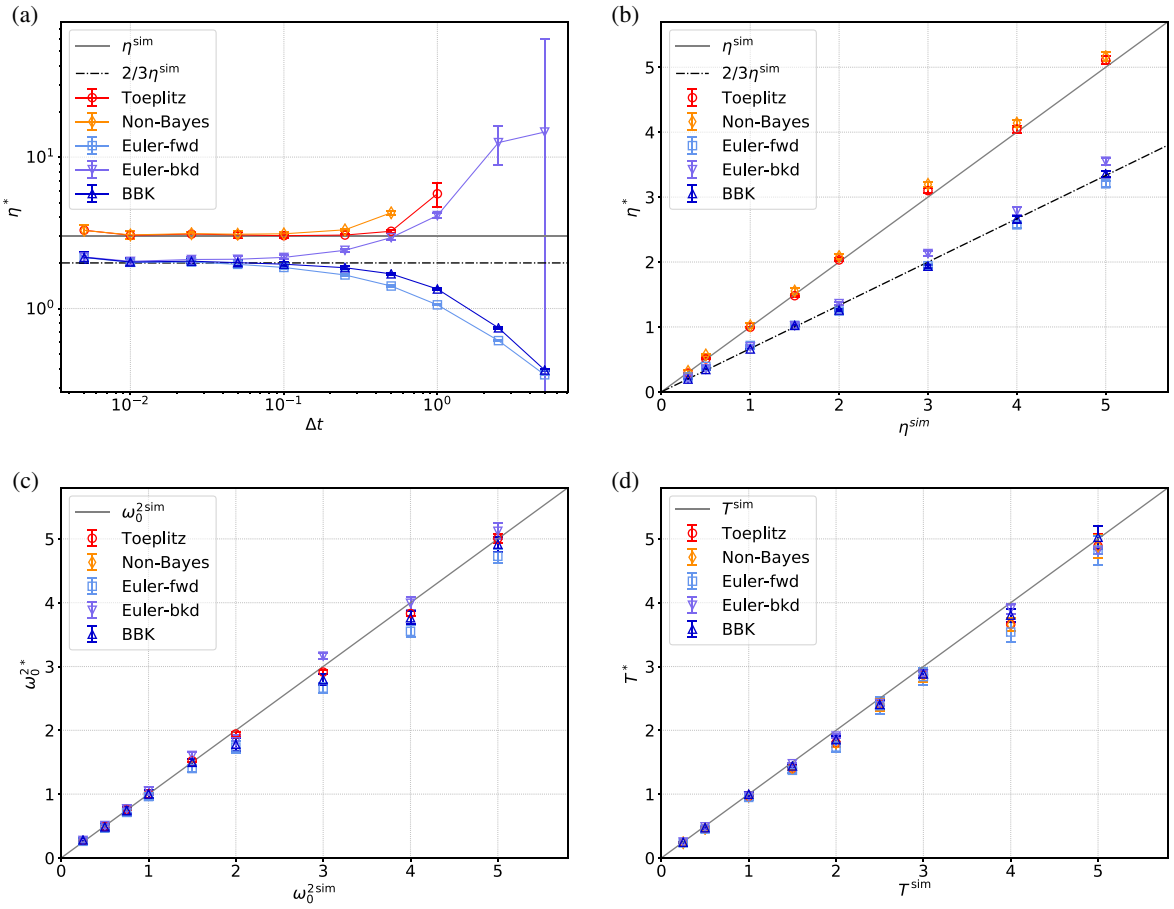

FIG. 1. Inference results for the stochastic harmonic oscillator. Sample trajectories are obtained from exact numerical integration of the set of first-order equations with parameters $\eta^{\rm sim}$, $\omega_0^{\rm sim}$, and $T^{\rm sim}$. The simulation time step $\tau^{\rm sim}$ is always equal to 0.005, and it corresponds to the minimum displayed value of Δt in (a). Points at higher values of Δt are obtained by applying the inference procedure to subtrajectories extracted from the original one. Each of the points displayed in (b)–(d) is obtained as a weighted average of the inference results for different Δt values in the range where the small Δt approximation is valid. Weights correspond to the squared inverse of the error bars, displayed in (a) for the η parameter. We compare the accuracy of all the schemes derived in Appendix B from a first-order Taylor-Itô expansion (Euler-fwd, Euler-bkd, and BBK) and from a second-order expansion (Toeplitz and non-Bayes). (a) Inferred values for the damping coefficient of the harmonic oscillator, η^*. Averages over ten sample trajectories of 5000 points (for any Δt) are reported with their 0.95 CI. Simulation parameters: $T = 1$, $\omega_0 = 1$, and $\eta = 3$. (b) Inferred damping coefficient η^* vs true simulation parameter $\eta^{\rm sim}$: Results from higher-order methods follow the line of slope 1, whereas numerical results from naïve methods fall on the line of slope $2/3$. The remaining parameters are fixed: $T = 1$ and $\omega_0 = 1$. (c) Inferred squared frequency of the harmonic oscillator ω_0^{2*} vs true simulation parameter $\omega_0^{2\rm sim}$. All the schemes give correct results in this case in the whole explored range of values. Simulation parameters: $\eta = 3$ and $T = 1$. (d) Inferred temperature T^* vs the true value of the simulation parameter $T^{\rm sim}$: Again, results from all schemes fall on the line of slope 1 in the whole explored range of values. Remaining simulation parameters: $\eta = 1.5$ and $J = 1$.

real experiments, where the time resolution is fixed by the acquisition apparatus, while the true microscopic timescale of the dynamics is unknown. Filtering the synthetic trajectories in time is a good blind inspection tool to check the robustness of the continuous description given by the inferred parameters, without prior knowledge about the timescales of the process. Moreover, this test on numerical simulations can help us identify the time window in which any dynamical inference scheme is expected to work: In discretizing the equations of motion, the implicit assumption is that Δt must be much smaller than the typical timescales of the process (η^{-1} and ω_0^{-1} in this example).

Results, reported in Fig. 1, show that a systematic error in the estimation of the damping coefficient emerges, which can be cast into a constant rescaling factor close to $2/3$ for the inferred value η^* as compared to the true value $\eta^{\rm sim}$. It is worth remarking that this rescaling is *independent of* Δt, as clearly visible in Fig. 1(a), so increasing the resolution of the acquisition instruments is of no help in improving the estimation of the damping coefficient. The same problem also occurs when using other variants of the EM scheme obtained from a Taylor-Itô expansion of the same order, as we illustrate in Appendix B 1. On the contrary, the estimation of the remaining parameters is in agreement

with the parameter values used in the simulations, as shown in Figs. 1(c)–1(d).

Numerical evidence for the stochastic harmonic oscillator agrees with the results of Refs. [12,14], who point out, in a non-Bayesian framework, the failure of the same naïve embedding strategy for second-order SDEs. We stress that the EM discretization is the simplest and most commonly used extrapolation of the derivative of an observed variable from its finite increment. This approximated estimation of the velocity works if one observes the system in the overdamped regime, i.e., when $\eta\Delta t \gg 1$ and $\omega_0/\eta < \infty$, and the effective dynamics can be described by a first-order equation. In this case, EM-based inference schemes provide in effect excellent results [21,22]. However, when a non-Markovian signal is observed, such as the partial observation of a higher-dimensional Markovian process, these schemes are bound to fail.

A simple argument can help us to understand what is missing and why the parameter η is the one affected by the approximation. Assuming that experimental averages perfectly reproduce ensemble averages, we can replace into Eq. (14) the known analytical expression for the self-correlation of the harmonic oscillator in the stationary regime $C(0)$, $C(\Delta t)$, and $C(2\Delta t)$. Since the underlying assumption of the whole procedure is that the time lag Δt between subsequent points is small, compared to the typical timescales of the dynamics, we can perform a Taylor expansion around $t = 0$, obtaining from Eq. (14) an expression for η^* depending only on the derivatives of $C(t)$ at $t = 0$:

$$\eta^* \simeq \frac{1}{\Delta t}\frac{2\dot{C}(0) - \frac{2}{3}\dddot{C}(0)\Delta t^2 - \frac{\ddot{C}(0)\dot{C}(0)}{C(0)}\Delta t^3}{2\dot{C}(0) + \ddot{C}(0)\Delta t + \frac{1}{C(0)}[\dot{C}(0) + \frac{1}{2}\ddot{C}(0)\Delta t]^2\Delta t}. \tag{15}$$

Knowing explicitly $C(t)$ for the harmonic oscillator [also reported in Appendix B 1, Eq. (B18)], one can compute the desired derivatives:

$$C(0) = \frac{T}{\omega_0^2}; \quad \dot{C}(0) = 0; \quad \ddot{C}(0) = -T; \quad \dddot{C}(0) = \eta T. \tag{16}$$

Proper combinations of these quantities allow us to extrapolate all the parameters of the model. The importance of the first derivative as a quantity to discriminate between first- and second-order dynamics in oscillatorlike models is stressed in Refs. [33,34], with explicit reference to complex interacting systems. Our point is that we can go beyond the binary answer provided by $\dot{C}(0)/C(0)$, proportional—through a timescale factor—to 1 or to 0 for first- or second-order dynamics, respectively, and give a quantitative estimation of the damping regime in which a system operates, employing all the derivatives at $t = 0$ up to the third one.

By replacing Eqs. (16) into Eq. (15), we obtain

$$\eta^* = -\frac{2}{3}\frac{\dddot{C}(0)}{\ddot{C}(0)}[1 + O(\Delta t)] = \frac{2}{3}\eta + O(\Delta t). \tag{17}$$

We find then, at the leading order, a rescaling factor of $2/3$, as observed in numerical tests. No rescaling factors appear for the other inferred parameters: Performing the same replacement and expansion of the analytical correlation functions in the inference formulas of T and ω_0, we see that the temperature and pulsation are correctly retrieved from proper combinations of $C(0)$ and $\ddot{C}(0)$.

This result gives us a clue to understand the origin of the Δt-independent rescaling factor for η. Looking back at Eq. (8), one realizes from simple dimensional analysis that the elimination of the velocity variables makes $O(\Delta t^{3/2})$ terms appear, even if the starting accuracy of the expansion is $O(\Delta t)$. This inconsistency means that Eq. (13) has been derived retaining only some of the $O(\Delta t^{3/2})$ contributions; in turn, the neglected ones produce missing $O(\Delta t^3)$ contributions to the fluctuations of x. This fact explains why Eq. (14) is incorrect and shows the need of higher-order discretization schemes for stochastic second-order dynamics.

We finally remark that this $2/3$ rescaling factor is not a specific feature of the stochastic harmonic oscillator but a recurrent trait in stochastic models of the form of Eq. (1). As rigorously proven by Gloter, the so-called quadratic variation of the discretized velocities (corresponding to an empirical estimate of the squared acceleration) uniformly converges to the expected value for the quadratic variation of the real unobserved velocities rescaled by $2/3$ [26]. These quadratic variations are $O(\Delta t^3)$, and the former one is the only directly measurable quantity containing the necessary dynamical information to disentangle the contribution of dissipation from diffusion and infer η in our setting [35].

C. Higher-order inference schemes

The lowest order of convergence required to develop any reasonable dynamical maximum-likelihood scheme is $O(\Delta t^{3/2})$. Since the mean square convergence of the infinitesimal increment of the process is what determines its statistical properties at any time, the minimum requirement for an inference method exploiting only local dynamical information is to reproduce fluctuations correctly at the leading order in Δt.

Independently of the details of the discretization, following the procedure outlined in Sec. II A, with $O(\Delta t^{3/2})$ accuracy, one reduces to a sequence of intertwined Gaussian integrals for the marginalization of $v_1 \ldots v_L$, which may be cumbersome to compute for an arbitrary length of the trajectory. Therefore, it is convenient to work again with update equations in x space. They can be obtained either from a temporal discretization of the

GLE (7) or from the elimination of the velocity variables in the discrete-time equations resulting from a second-order Taylor-Itô expansion of the Markov process in Eq. (3). In the first case, since the same exponentially decaying kernel propagates both the noise and the initial condition in Eq. (7), it is possible to manipulate the integrated GLE to find a stochastic difference equation that does not contain v_0 and is driven by a short correlated effective noise:

$$x_{n+1} - x_n - e^{-\eta\Delta t}(x_n - x_{n-1}) = \frac{1-e^{-\eta\Delta t}}{\eta}\int_{t_{n-1}}^{t_{n+1}} \Psi(t-t_n) f(x(t)) dt + \zeta_n, \tag{18}$$

where

$$\zeta_n = \int_{t_{n-1}}^{t_{n+1}} \Psi(t-t_n)\xi(t)dt; \tag{19}$$

$$\Psi(t) = \begin{cases} \frac{e^{\eta t}-e^{-\eta\Delta t}}{1-e^{-\eta\Delta t}} & \text{if } -\Delta t < t < 0; \\ \frac{1-e^{\eta(t-\Delta t)}}{1-e^{-\eta\Delta t}} & \text{if } 0 < t < \Delta t. \end{cases} \tag{20}$$

Correspondingly, the *first discretize, then marginalize* strategy provides a stochastic difference equation with the same properties. We detail both procedures in Appendix A.

Concentrating on the case of the stochastic harmonic oscillator, any consistent discrete-time description in x space takes the form of a linear stochastic difference equation like

$$x_{n+1} + \alpha x_n + \beta x_{n-1} = \zeta_n, \tag{21}$$

where the inhomogeneous terms ζ_n are still Gaussian random variables of null mean, but they are no longer independent. This feature is the crucial difference with the Euler-Maruyama scheme, which takes into account only the diagonal entries of the covariance matrix $C_{nm} = \langle \zeta_n \zeta_m \rangle$.

Equation (21) defines an affine map:

$$\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_{L-1})^\top \mapsto \mathbf{x} = (x_2, \ldots, x_L)^\top = \mathbf{M}^{-1}\boldsymbol{\zeta} + \mathbf{x}_0, \tag{22}$$

where $M_{ij} = \delta_{i,j} + \alpha\delta_{i,j-1} + \beta\delta_{i,j-2}$ and $\mathbf{x}_0 = (x_0, x_1, 0, \ldots, 0)^\top$, which can be generalized to a nonlinear transformation when anharmonic forces are present. This map can be exploited, when the covariance matrix $\mathbf{C}$ and its inverse are known, to write the new, higher-order, dynamical likelihood. For the harmonic oscillator, it reads

$$P_{(2)}(x_L, \ldots, x_2 | x_1, x_0) = \frac{1}{Z}\exp -\frac{1}{2}\sum_{n,m=1}^{L-1}(x_{n+1} + \alpha x_n + \beta x_{n-1}) C^{-1}{}_{nm}(x_{m+1} + \alpha x_m + \beta x_{m-1}), \tag{23}$$

where Z is the normalization constant:

$$Z = [(2\pi)^{L-1}\det \mathbf{C}]^{1/2} = \left[\prod_{k=1}^{L-1} 2\pi\lambda_k\right]^{1/2}, \tag{24}$$

with λ_k the kth eigenvalue of the covariance matrix $\mathbf{C}$. The effective parameters α and β, as well as the entries of the covariance matrix, are known combinations of the parameters of the model, whose details depend on the adopted discretization scheme. In the following results, we adopt $\alpha = -1 - e^{-\eta\Delta t} + \omega_0^2\Delta t(1-e^{-\eta\Delta t})/\eta$ and $\beta = e^{-\eta\Delta t}$.

For well-chosen α, β, and C_{nm}, Eqs. (21) and (23) are exact, in the limit $L \to \infty$. Thanks to linearity, it is possible to design an exact integration algorithm for the Markov process (3) at any time-step increment Δt [32]. For nonlinear generalizations of $f(x)$, the exact Gaussian character of the random increment is lost. However, at leading order in Δt, a multivariate Gaussian distribution still represents a good approximation for the distribution of the random increments ζ_n appearing in the x update equation, which takes the form

$$x_{n+1} + F(x_n, x_{n-1}; \boldsymbol{\mu}) = \zeta_n, \tag{25}$$

with $\boldsymbol{\mu}$ a set of effective parameters. The corresponding generalization of Eq. (23) can be obtained (see Appendix B 2).

To the order Δt^3, for both linear and nonlinear second-order processes, one can deduce from Eq. (19) that $\mathbf{C}$ has a "nearest-neighbor" structure of the kind

$$C_{nm} = \langle \zeta_n \zeta_m \rangle = a\delta_{n,m} + b\delta_{n,m\pm1}, \tag{26}$$

where

$$a \simeq \frac{2}{3}2T\eta\Delta t^3; \qquad b \simeq \frac{1}{6}2T\eta\Delta t^3. \tag{27}$$

Hence, the covariance matrix has the form of a symmetric $(L-1)\times(L-1)$ tridiagonal Toeplitz matrix. These mathematical features carry a deep physical meaning: First of all, the presence of nonvanishing off-diagonal elements is the signature of a colored noise. Second, the fact that the matrix is banded means that the correlation of the noise variables is short ranged, i.e., that the associated memory

kernel, in a continuous-time description, decays fast [9]. Finally, the Toeplitz structure is synonymous with shift invariance.

A more careful derivation of the update equations in x space would require shift invariance not to hold and the first entry of the covariance matrix C_{11} to be different from the other elements of the main diagonal. Equation (21) is, in fact, not valid for the first integration step, where the initial conditions intervene. In this respect, the structure of the data also poses the problem of the elimination of the initial condition v_0 in favor of x_0 and x_1. Even if not able to perform it explicitly without stationarity assumptions, we can argue (see Appendix A) that it has the effect of modifying the covariance matrix in the following way:

$$C = \begin{pmatrix} \tilde{a} & b & \cdots & 0 \\ b & a & \cdot & \vdots \\ \vdots & \cdot & \ddots & b \\ 0 & \cdots & b & a \end{pmatrix}, \tag{28}$$

where the shift invariance expressed by the Toeplitz structure of Eq. (26) is then broken at the beginning of the time series. Despite that, the error we make by replacing $\tilde{a}$ with a in the quasi-Toeplitz matrix (28) is negligible in the limit of long trajectories, as discussed in Sec. IV and checked in Fig. 4. Intuitively, since the breaking of the shift invariance occurs only at the first step, the longer the trajectory, the more similar this is to a truly shift invariant situation. Notice that what matters is not the total length $(L+1)\Delta t$ of the trajectory in units of the physical time scales of the process, but just the number of points $L+1$ of which the trajectory is made up [36].

Apart from the difficulty in determining correctly $\tilde{a}$, the advantage of replacing the true covariance matrix Eq. (28) with a Toeplitz matrix is that the inverse of the Toeplitz matrix is explicitly known, as well as the eigenvalues [37,38]:

$$C^{-1}{}_{nm} = \frac{2}{L}\sum_{k=1}^{L-1} \frac{\sin(\frac{nk\pi}{L})\sin(\frac{mk\pi}{L})}{a + 2b\cos(\frac{k\pi}{L})}; \tag{29}$$

$$\lambda_k = a + 2b\cos\left(\frac{k\pi}{L}\right). \tag{30}$$

Let us highlight that the inverse of the covariance matrix does not preserve a banded structure, meaning that, even if noise correlations are local in time, two-time functions of every pair of points of the trajectory enter into the minus-log-likelihood. Hence, Eq. (23) cannot be factorized. Factorization corresponds to a block structure for $\mathbf{C}^{-1}$, which implies a block structure for $\mathbf{C}$. This property is incompatible with the tridiagonal Toeplitz or quasi-Toeplitz nature of the covariance matrix, where off-diagonal elements are of the same order as the diagonal ones.

Nonetheless, having built an explicit discrete path integral measure, a maximum-likelihood approach is practicable, and it reduces to minimizing the quantity $\mathcal{L} = -\ln P(x_L, \ldots, x_2|x_1, x_0)$ with respect to the parameters of the model. Thanks to the regularities of Eq. (23), the minimization of $\mathcal{L}$ can be performed analytically in the case of the harmonic oscillator and, in general, of simple single-particle systems. The optimization procedure can be performed semianalytically also for many-particle systems, like active agent-based microscopic models or spatially discrete counterparts of field theoretical models. In these cases, an additional parameter is typically the interaction range of effective pairwise potentials, which may depend on a different (measurable) variable than the fieldlike observable x. In general, once an expression for $\mathcal{L}$ is given, a large number of optimization algorithms are available to minimize it with respect to all the extra parameters that do not allow for a full analytical approach.

Complete inference formulas for one-dimensional harmonic and anharmonic oscillators and for a system of many coupled harmonic oscillators with a parameter-dependent connectivity matrix are reported in Appendixes B 2 and B 3. In all cases, optimal parameter values are given by combinations of all the two-time functions up to the length of the trajectory and not only those computed at a temporal distance of zero, one, and two time steps.

For the noninteracting case, we test the developed schemes numerically by applying the inference formulas to synthetic stochastic trajectories of two reference processes: the Brownian motion in a harmonic potential and the Brownian motion in a symmetric anharmonic potential $V(x) = \frac{1}{2}kx^2 + \frac{1}{4}\lambda x^4$. The equations of motion corresponding to the latter read

$$\dot{x} = v, \qquad \dot{v} = -\eta v - kx - \lambda x^3 + \sigma\xi, \tag{31}$$

where we choose a unitary mass particle, $\sigma^2 = 2T\eta$, and $\xi(t)$ as a white noise. We generate synthetic trajectories as in Ref. [39] and subsample them by progressively increasing the time separation Δt between subsequent observed points.

The comparison with naïve inference schemes for the example of the harmonic oscillator confirms the analytical predictions (Fig. 1). In any damping regime, the higher-order inference method outperforms the naïve scheme in two ways: perturbatively, since the convergence of the parameter estimators is extended to a larger Δt window due to the higher-order Taylor-Itô expansion [an example in Fig. 1(a)], and nonperturbatively in Δt, since no rescaling factor for the η parameter is required [Fig. 1(b)]. The different behavior of the various schemes at large Δt, where the series expansion is nonasymptotic, is probably related to the details of the discretization rules and their stability properties.

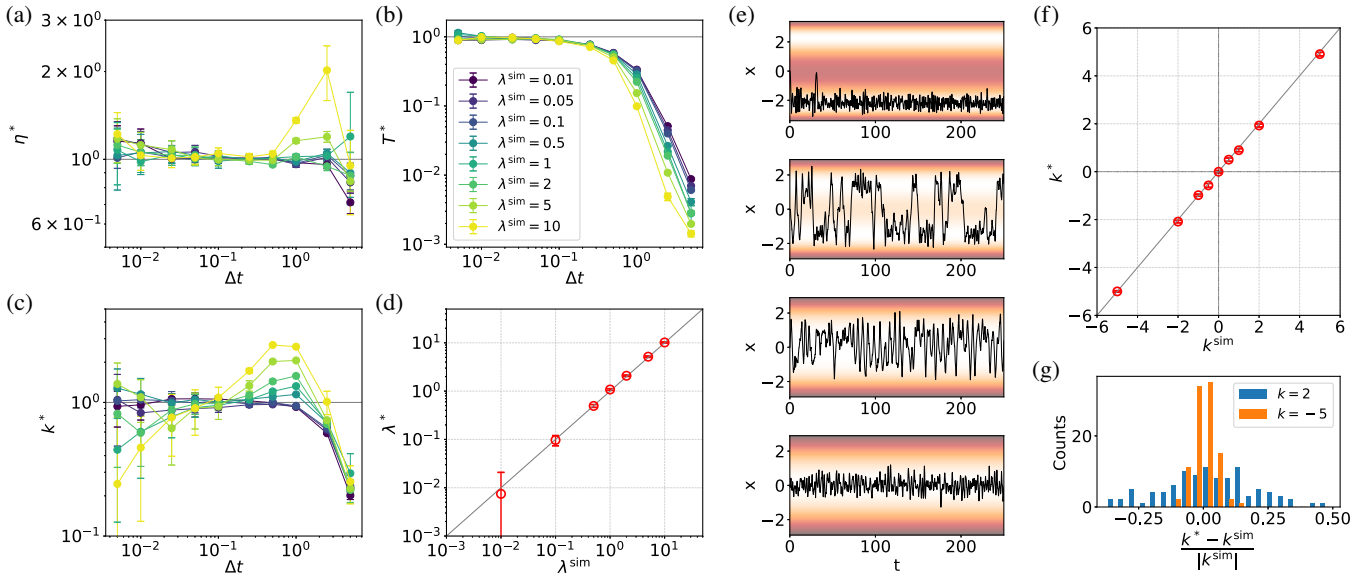

FIG. 2. Bayesian inference of the dynamical parameters of a Brownian motion in a force field $f(x) = V'(x)$, with $V(x) = \frac{1}{2}kx^2 + \frac{1}{4}\lambda x^4$. Only the Toeplitz method is applied; as for the harmonic oscillator, in (a)–(d) and (f), ten sample trajectories of length 5000 points are considered for each Δt. Error bars are 0.95 CI. (a)–(c) Inferred model parameters against subsampling parameter Δt. The true value is equal to 1 in all cases and is marked by the straight gray line. (d) Inferred vs true value of the nonlinear coefficient λ. (e) Excerpts of sample trajectories in various landscapes. The strength of the confining potential is qualitatively indicated by the color map, with light areas corresponding to the minimum of the potential. The following parameters of the simulation are kept fixed: $T = 1$, $\eta = 1$, and $\lambda = 1$. By varying the parameter k, we realize, from top to bottom, a strong confinement in a double-well potential, with long exit times, at $k = -5$; a switching dynamics with relatively short switching times, at $k = -2$; a marginal situation at $k = 0$; and confined Brownian motion in the vicinity of the origin at positive values of k ($k = 5$). (f) Inferred vs true value of the parameter of the linear force k, assuming both positive and negative values. (g) Histogram of counts for the relative distance of the inferred parameter k^* to the simulation parameter $k^{\rm sim}$. With fixed $\lambda^{\rm sim} = 1$ and $k^{\rm sim} = \{2, -5\}$, the weight of anharmonicity varies, but the variance of all the estimated parameters seems to be unaffected. As a result, relative errors decrease for larger $|k|$. 100 trajectories are sampled for each k value shown in the histogram, and $\Delta t = 0.025$ in all cases.

Figure 2 shows numerical results based on the Toeplitz inference scheme for the anharmonic stochastic oscillator for varying values of the parameters $\lambda^{\rm sim}$ [Figs. 2(a)–2(d)] and $k^{\rm sim}$ [Figs. 2(e)–2(g)]. In all the explored regimes, the inference scheme provides excellent results, showing, in particular, that no bias is introduced by the possible imbalance between linear and nonlinear force terms [values close to the origin are correctly estimated in Figs. 2(d) and 2(f)], even if, for a fixed Δt, an increase in the relative error or more noisy estimations cannot be prevented in these conditions [Fig. 2(g)]. Moreover, no bias is introduced by the fact that, when k assumes a negative value, the particle may be confined in a single minimum of the double-well potential for all the length of the sampled trajectory [see Fig. 2(e)].

D. Generalization to multiplicative noise

In order to understand the limits and full potential of the method, we focus in this section on possible generalizations to the case of nonadditive noise. An adaptation of our non-Markovian Bayesian inference scheme can be developed for the following class of multiplicative processes:

$$\ddot{x} = -\eta\dot{x} + f(x) + \sigma(x)\xi, \tag{32}$$

with $\xi(t)$ a standard white noise and initial conditions $x(0) = x_0$, $\dot{x}(0) = v_0$. This model has two features: linear dissipation and a velocity-independent diffusion coefficient proportional only to $\sigma^2(x)$. Under these conditions, the memory kernel of the GLE associated to Eq. (32) is explicitly known, and, following the same procedure that leads to the discretization of the additive process in Sec. II C, we obtain an approximated discrete-time update rule of the form

$$x_{n+1} - x_n - e^{-\eta\Delta t}(x_n - x_{n-1}) - \frac{1 - e^{-\eta\Delta t}}{\eta}\Delta t f(x_n) = \zeta_n, \tag{33}$$

where the stochastic term is defined as

$$\zeta_n = \frac{1 - e^{-\eta\Delta t}}{\eta}\int_{t_{n-1}}^{t_{n+1}} dt'\Psi(t' - t_n)\sigma(x(t'))\xi(t'). \tag{34}$$

The function $\Psi(t)$ is defined in the same way as in Eq. (20).

The approximation of the force term in Eq. (33) corresponds to that of the Langevin impulse integrator [40]. Alternative numerical integration schemes for GLEs, such as the stochastic Verlet algorithm [41], can also be used.

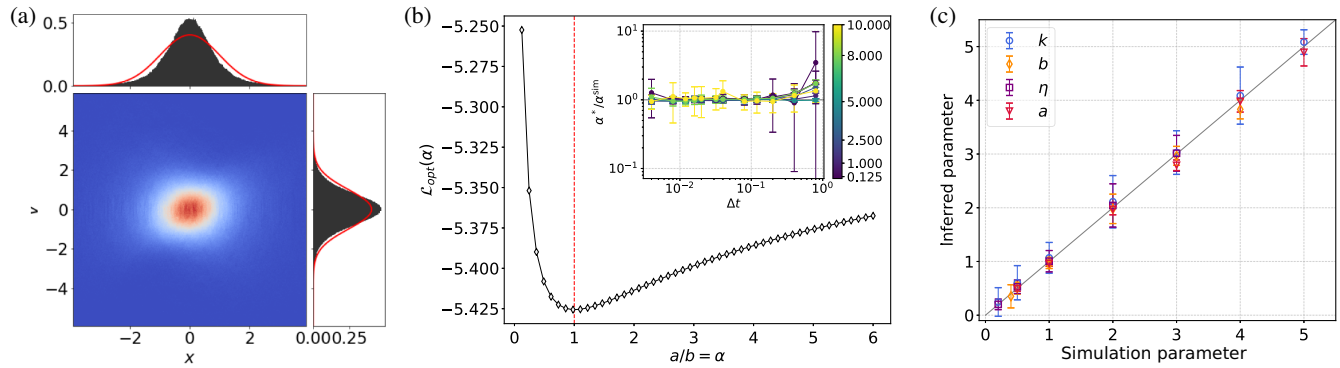

FIG. 3. Inference method applied to a multiplicative process. The process is described by Eq. (32) with $f(x) = -kx$ and $\sigma(x) = \sqrt{a + bx^2}$. (a) The fraction of time spent by the system in each region of the phase space for a sample trajectory of length 4×10^4, with $k = 1$, $\eta = 1$, $a = 1$, $b = 1$, and initial condition $(x_0 = 0, v_0 = 0)$. There is a clear difference with the Gaussian distribution having the same second moment (red line), showing the effect of the multiplicative noise. (b) Analytically optimized negative log-likelihood as a function of the effective parameter $\alpha = a/b$, computed on a sample subtrajectory of 5000 points, $\Delta t = 0.016$, with the same parameters as in (a). In the inset, optimal values of α as a function of Δt are reported. Error bars correspond to 0.95 CI on ten sample trajectories of 5000 points for each Δt. The color code refers to the value of α^{sim}, measuring the relative contribution of additive and multiplicative part of the noise term. (c) Performance of the method in inferring the whole set of parameters of the model.

From now on, we implicitly refer to the Itô integration prescription. However, due to the fact that $\sigma(x)$ depends only on the configurational degree of freedom x, the mean square convergence of ζ_n is not affected by a switch to the Stratonovich convention. As a result, one can say that, up to $O(\Delta t^3)$, stochastic terms satisfy

$$\langle \zeta_n \zeta_m \rangle \simeq \frac{2}{3}\Delta t^3 \sigma^2(x_n)\delta_{n,m} + \frac{1}{6}\Delta t^3 \sigma(x_n)\sigma(x_m)\delta_{n,m\pm 1}. \quad (35)$$

This choice of off-diagonal terms ensures the positiveness of the matrix, if $\sigma(x) > 0$ [42]. The covariance matrix also preserves a tridiagonal symmetric structure. However, the Toeplitz property is lost, since, in the presence of multiplicative noise, shift invariance cannot hold. Nevertheless, we can build an efficient maximum-likelihood inference routine. Let us rewrite the minus-log-likelihood associated to Eq. (33) as

$$\mathcal{L} = \frac{1}{2}\sum_{k=1}^{L-1} \ln \lambda_k(\mathbf{x};\boldsymbol{\nu}) + \sum_{n,m=1}^{L-1} [x_{n+1} + F(x_n, x_{n-1};\boldsymbol{\mu})] C^{-1}{}_{nm}(\mathbf{x};\boldsymbol{\nu})[x_{m+1} + F(x_m, x_{m-1};\boldsymbol{\mu})], \quad (36)$$

so that we can distinguish between the subset of parameters $\boldsymbol{\mu}$, including η and the parameters of the conservative potential, and the subset ν appearing in the x-dependent diffusion coefficient $\sigma(x;\boldsymbol{\nu})$. For the parameters in the former set, analytical formulas for their max-likelihood estimators can be found as functions of $\boldsymbol{\nu}$, while the latter generally requires numerical optimization [unless $\sigma(\mathbf{x};\boldsymbol{\nu})$ is univariate and has a purely multiplicative dependence on its single parameter]. The effective cost function can be evaluated, also in the case of long trajectories, once the inverse and the spectrum of the symmetric tridiagonal matrix are computed.

To illustrate the method, we apply it to the multiplicative process in Eq. (32), with $f(x) = -kx$ and $\sigma(x) = \sqrt{a + bx^2}$, where a and b are non-negative parameters. In this case, the max-likelihood procedure can be reduced to a one-dimensional numerical optimization. Complete inference formulas are reported in Appendix B 4, and the results are shown in Fig. 3. These results confirm that the method provides a reliable inference tool also in the case of a nonequilibrium multiplicative process, independently of the relative strength of the additive and multiplicative contributions to the noise term, and that the procedure does not require equilibrium assumptions to work, nor does it exploit the fluctuation dissipation theorem.

III. ALTERNATIVE NON-BAYESIAN APPROACH

Alternative inference approaches to the maximum-likelihood method are also possible. Several examples are known in the literature: The most general ones, applicable to a vast class of second-order stochastic processes, derive the parameters of the assumed model (in the form of a SDE or of a chosen set of projection functions) through a fitting procedure on measurable quantities, typically involving conditional moments of the increments of the process

[11–14]. Even in this case, the method is local: The relations used for fitting can be found through a Taylor-Itô expansion even when a nonlocal solution in time is unknown. Other strategies have been proposed with a reformulation of the task—having relevant applications in chemical physics and molecular dynamics—i.e., not to learn the best model for the measured variables but to find from higher-dimensional data the coarse-grained dynamics of a given system [43,44].

In this section, we put ourselves in a simpler framework than that of Refs. [11–13] and derive non-Bayesian parameter estimators just for the stochastic harmonic oscillator, in order to compare on this example the non-Bayesian methodology and our maximum-likelihood dynamical inference scheme. From update equations in position space like Eq. (21), obtained from an $O(\Delta t^{3/2})$ Taylor-Itô expansion, some relations between experimental correlation functions and model parameters can be found. Let us take the update equation of the Langevin impulse integrator [40]:

$$x_{n+1} = x_n + e^{-\eta\Delta t}(x_n - x_{n-1}) + \frac{1 - e^{-\eta\Delta t}}{\eta}\omega_0^2 \Delta t x_n + \zeta_n \tag{37}$$

with ζ_n the Gaussian random variables characterized by Eq. (26). Multiplying both sides of Eq. (37) by x_m, for $m \in \{n-1, n, n+1\}$, and self-consistently averaging over the noise distribution yields a set of three independent equations, from which all the parameters of the dynamical model can be extracted (explicit formulas are derived in Appendix B 5).

Notice that, in contrast to the max-likelihood inference method, the obtained relations can involve only three types of temporal correlation functions: equal-time, one-time-step, and two-time-step correlations. Even if we are not using all the exploitable information carried by an N-point trajectory (the operation outlined above could, in principle, be performed for all x_m), this one is the optimal minimal choice. Indeed, the shape of the temporal correlation function at small times contains substantial dynamical information. Moreover, due to the finite length of the trajectories, two-time quantities, like correlation functions, are typically better estimated at small time differences than at large ones.

As expected, parameter estimators provide good values without rescaling. Unfortunately, however, we cannot extend this approach to interacting systems, where an interaction range is needed to parametrize the potential. As these formulas do not come from the optimization of any cost function, there is no efficient numerical strategy to find the best parameters of the interaction potential. The problem is bypassed if no assumption is made about the structure of the interaction, and a different parameter is associated to each element pair in the system. In this framework, however, severe overfitting issues may emerge as well as numerical scaling problems, since the number of parameters grows roughly quadratically with the system size. We remark that this scaling curse does not afflict all non-Bayesian inference methods [11] but only the simple one used here to compare its results with our Bayesian scheme.

Finally, it is important to specify the probability density function with respect to which we are taking the averages in Eq. (37). Since, in order to compute $\langle x_n \xi_n \rangle$ and $\langle x_{n+1} \xi_n \rangle$, we self-consistently use the same update rule and the same shift-invariant noise statistics, we argue that we implicitly introduce a stationarity assumption, overcoming the problem anticipated in Sec. II A and better discussed in Sec. IV. As a result, the inference formulas obtained in this way do not require any rescaling factor, for any length of the trajectory.

IV. ROLE OF THE UNOBSERVED INITIAL CONDITION

Once colored noise is included to take into account the non-Markovian character of the partially observed process, the remaining problem in the application of the Bayesian methodology to second-order stochastic models lies in the elimination of the unobserved initial condition. To explain this problem, let us take a step back.

In a maximum-likelihood setting, the first task is to calculate the probability of observing a given sequence of data points, knowing the parameters of the model $\boldsymbol{\lambda}$. In first-order stochastic processes, when all the degrees of freedom allowing for a Markovian description of the dynamics are experimentally accessible, there is no ambiguity on how this likelihood should be computed (see, for example, Ref. [21]). On the other hand, for second-order stochastic processes, the inference problem may turn out to be ill defined. For a first-order model

$$\dot{y} = g(y) + \xi, \tag{38}$$

with initial condition $y(0) = y_0$, the propagator is defined as $P(y_L, \ldots, y_1 | \boldsymbol{\lambda}; y_0)$. Here, y_0 represents quantities that do not change in the inference procedure and the same in the posterior, the likelihood, and the prior: the initial condition and the structure of the model. We introduce a semicolon to separate the quantities that are not updated in the inference. For a second-order stochastic process, the initial condition is given by the pair $x(0) = x_0$, $\dot{x}(0) = v_0$, and the propagator is $P(x_L, \ldots, x_1 | \boldsymbol{\lambda}, x_0, v_0)$. However, unlike x_0, the initial condition on the velocity is not empirically known, so the propagator does not result in a likelihood of the form of $P(x_L, \ldots, x_1 | \boldsymbol{\lambda}; x_0, v_0)$.

Let us briefly note that this problem is strictly connected to the embedding problem in stochastic processes and that the only consistent way to bypass it is to use the steady state distribution of v_0. Nonetheless, in Sec. II C, we decide to deal with the initial condition problem in a different way. First, the choice of the discretization scheme confines the

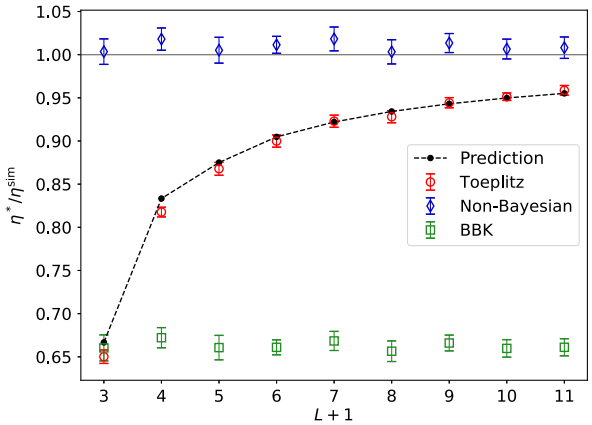

FIG. 4. Numerical validation of the finite-size distortion introduced by the shift-invariant approximation. Black points connected by dashed lines represent the analytical prediction about the rescaling factor $\varrho(L+1) = \eta^*(L+1)/\eta^{\rm sim}$, with $\varrho(L=3) = 2/3$ (shortest possible trajectory) and $\varrho(L+1) \to 1$ monotonically as $L \to \infty$. Numerical results are in agreement with this prediction. As expected, no dependence on the trajectory length is found for the non-Bayesian method nor for Euler-like methods (BBK used here—see Appendix B 1 for details).

initial condition problem only to the first time step, independently of the total number of data points and the relation of the decay time of the memory kernel to Δt. Neglecting the breaking of shift invariance, we introduce a Toeplitz approximation for the noise covariance matrix: This approximation works well for long trajectories (with many data points), whereas it fails for very short ones. The convergence is, however, quite fast, as shown in Fig. 4. The advantage of this strategy is twofold: It is simpler than exact marginalization and applies even when a steady state distribution is not available (e.g., in the multiplicative case).

Remarkably, the problem of the elimination of the initial condition on the first derivative of the observed variable does not affect the non-Bayesian approach. As a result, non-Bayesian methods apply even to (multiple) disconnected triplets of points or, in general, to disconnected small sequences, if a fragmented observation of the system is the only one achievable. On the contrary, the Toeplitz method is exact only in the infinite trajectory limit, so the smaller the number of subsequent points, the less accurate the inference scheme becomes. In other words, what matters in this case is not only the total number of points for statistical reasons—which is the only thing to worry about in all the other developed schemes—but also their succession in time.

We check this prediction in numerical simulations of the stochastic harmonic oscillator, keeping constant the total number of points used in the inference procedure, $(L+1)n_S$, and adapting the number of samples n_S as the length $L+1$ of the sample trajectories is varied. A significant deviation of the inferred value from the simulated one is visible in Fig. 4 for small values of L. For small L, it is also possible to approximately estimate the distortion introduced by the finite size of the trajectory under the Toeplitz assumption. Following the same idea that leads to the prediction of the $2/3$ function for the η parameter of the harmonic oscillator, one can expand the two time correlation functions appearing in the Toeplitz inference formulas for small L, obtaining

$$\eta^* \simeq -\frac{1}{\Delta t}\ln\left(1 + \varrho(L+1)\frac{\dddot{C}(0)}{\ddot{C}(0)}\Delta t\right)[1 + O(\Delta t)], \qquad (39)$$

from which we deduce that the Δt-independent rescaling factor of the damping coefficient can be identified with $\varrho(L+1)$ in Eq. (39). The first few values of these rescaling factors are $\varrho(3) = 2/3$, $\varrho(4) = 5/6$, $\varrho(5) = 7/8$, and $\varrho(6) = 19/21$, in good agreement with numerical results. The exact value is retrieved only in the $L \to \infty$ limit, yet time lapse recordings in common motility observation experiments are typically composed by a much larger number of frames than those shown in Fig. 4. Although we show that the wrong marginalization of the initial condition can play a role, in practice, this effect can hopefully be neglected in many situations.

V. INTERACTING CASE

Following our original objective to develop an inference strategy for natural flocks of birds, we generalize the inference equations of Sec. II and perform numerical simulations of the topological inertial spin model (ISM) on a nonevolving random lattice at low temperature. The model, introduced to account for experimentally observed features that could not be explained within the framework of first-order processes [6,45], represents a second-order generalization of the well-known Vicsek model. The stochastic equations of motion in three dimensions read

$$\dot{\mathbf{r}}_i = \mathbf{v}_i, \qquad \dot{\mathbf{v}}_i = -\frac{1}{\chi}\mathbf{v}_i \times \mathbf{s}_i,$$
$$\dot{\mathbf{s}}_i = -\frac{\eta}{\chi}\mathbf{s}_i + \frac{1}{v_0^2}\sum_j J_{ij}(\mathbf{v}_i \times \mathbf{v}_j) + \boldsymbol{\xi}_{i\perp}, \qquad (40)$$

where the indexes $i, j = 1, \ldots, N$ label different individuals, v_0 is the constant modulus of each velocity vector $\mathbf{v}_i$, and $\boldsymbol{\xi}_{i\perp}$ is the orthogonal projection to $\mathbf{v}_i$ of a three-dimensional white noise of parameters T and η: $\langle \boldsymbol{\xi}_{i\perp}(t) \cdot \boldsymbol{\xi}_{j\perp}(s)\rangle = 2\delta_{ij}2T\eta\delta(t-s)$. Motivated by the findings of Ref. [46], we choose to parametrize the coupling constant as $J_{ij} = Jn_{ij}$, where $n_{ij} = 1$ if bird j is among the first n_c nearest neighbors of bird i, whereas it takes a null value otherwise.

In the ordered phase, the spin-wave expansion of the equations of motion of the inertial spin model linearizes the force terms, and Eq. (40) takes the form of a set of SDEs for N coupled harmonic oscillators [21]:

$$\chi\ddot{\boldsymbol{\pi}}_i = -\eta\dot{\boldsymbol{\pi}}_i - J\sum_{j=1}^{N}\Lambda_{ij}\boldsymbol{\pi}_j + \tilde{\boldsymbol{\xi}}_{i\perp}, \quad i = 1, \ldots, N. \tag{41}$$

Here, $\boldsymbol{\pi}_i$ are the birds' normalized velocity fluctuations, lying on the orthogonal plane to the direction of collective motion, $\Lambda_{ij} = n_c\delta_{ij} - n_{ij}$ is the discrete Laplacian of the birds' network, and $\tilde{\boldsymbol{\xi}}_{i\perp}$ is now a two-dimensional white noise that lives on the same plane as $\boldsymbol{\pi}_i$. To leading order, it is described by the parameters T and η appearing in Eq. (40). For a full derivation of the equations of motion in the spin-wave approximation, see Appendix C. Thanks to the linearity of Eq. (41), the same inference strategy one can develop for a system of coupled harmonic oscillators applies also to the inertial spin model in the highly polarized phase.

For the sake of simplicity, in our simulations, we discard the first equation of Eq. (40) and keep the birds' reciprocal positions fixed. The dynamical maximum-likelihood approach, however, should work even when reshuffling birds' reciprocal positions and static approaches fail, since at each time step it is possible to reconstruct the neighborhood of each individual and compute the associated time-dependent observables [21,22,47]. Birds'

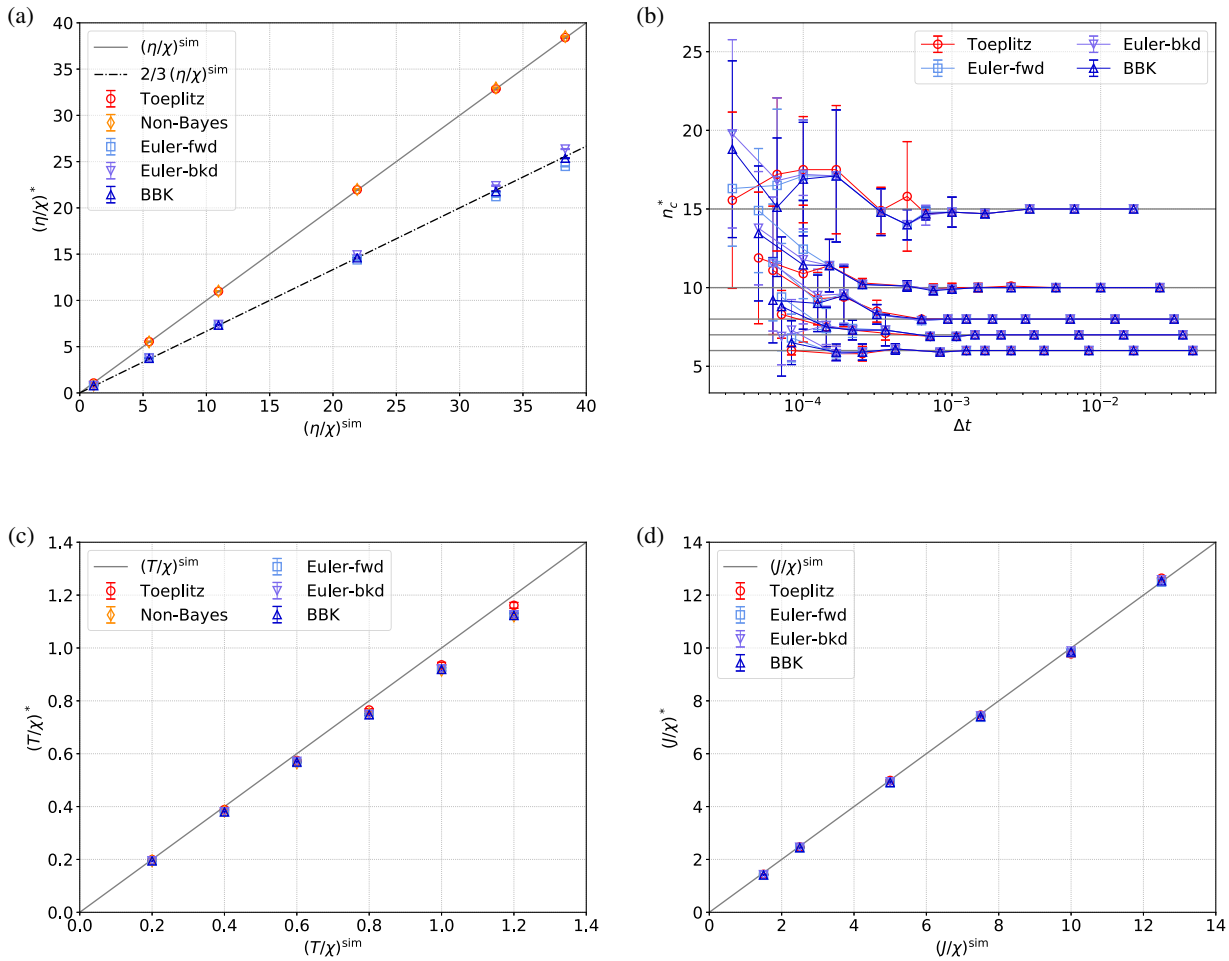

FIG. 5. Inference results for the inertial spin model. (a) Inferred values for the effective damping coefficient η/χ. We notice the emergence of a $2/3$ rescaling factor for naïve methods derived from first-order Taylor-Itô expansions. (b) Inferred topological interaction range from numerical minimization of the reduced minus-log-likelihood, which is properly defined only in the Toeplitz scheme and in the three lower-order variants of the Euler scheme. (c) Inferred values for the parameter T/χ, as derived from Eq. (41). One notices a slight divergence from the slope-1 line, which is especially evident at large temperatures. This divergence is due to the spin-wave approximation (SWA), whose first correction impacts only the temperature parameter and can be explicitly evaluated, as explained in Appendix C. (d) Inferred values of the interaction strength $(J/\chi)^*$ vs the parameter value used in simulations, $(J/\chi)^{\rm sim}$. All methods retrieve the correct results. We remark that only the parameters in the left panels, η/χ and T/χ, can be estimated by the non-Bayesian method. In all the simulations, we take flocks of $N = 1000$ birds. Additional information about the choice of the model parameters and numerical methods can be found in Appendix D. Points in (a), (c), and (d) are obtained as in the case of the harmonic oscillator (see Fig. 1). For the $O(\Delta t^{1/2})$ methods, we consider the different integration schemes defined in Appendix B 1.

reshuffling would introduce an effective nonlinearity which, like in the noninteracting case, is not supposed to modify the leading Gaussian nature of the propagator at small Δt.

We apply and compare different inference strategies to the synthetic trajectories. Results are in qualitative agreement with those of the harmonic oscillator. In particular, the expected rescaling factor of $2/3$ for the damping coefficient is retrieved using any EM-like scheme, as shown in Fig. 5(a). This fact corroborates that the emergence of this $2/3$ factor is a universal feature of second-order stochastic processes, coming from the interplay between the terms containing second- and first-order time derivatives, rather than the kind of forces which are applied to the system. Again, Bayesian and non-Bayesian inference schemes derived from a higher-order expansion do not require any rescaling—at least for sufficiently long trajectories.

As already mentioned, however, there are some relevant differences with respect to the simple noninteracting case. First of all, the additional difficulty we must face in the case of N-body dynamics is that of estimating the interaction range. Since an explicit analytical minimization of the minus-log-likelihood is not operable, a numerical approach is needed. The problem is, however, algorithmically tractable, since it simply consists of a one-dimensional optimization problem. Moreover, if the parametrization of the n_{ij} matrix discussed above is adopted, n_c is a discrete parameter, so the exact minimum value can always be found [see Fig. 5(b)]. Wrong estimations of the topological interaction range can be due to a blurred reconstruction of the likelihood from the data. As the number of birds N or the number of trajectory points L is increased, the improved statistics smoothens the rugged reconstructed likelihood, and the real minimum becomes easier to detect. To this end, another parameter playing a relevant role is the time lapse Δt: When the separation between subsequent data points is very small compared to the timescales of the system, increments are also very small. Smaller increments correspond to smaller quantities to minimize, which are then subject to bigger relative errors. This effect is at the origin of what we observe in Fig. 5(b).

Once the optimal value of n_c is recovered, it is then used to compute the spatially structured correlation functions which enter into the formulas of the remaining parameters. Non-Bayesian methods are not based on any likelihood definition and, as a result, do not allow us to infer n_c. Despite that, an approximated estimation of the effective temperature T/χ and of the damping coefficient η/χ is still possible, as shown in Figs. 5(a) and 5(c). On the contrary, the parameters associated to the interaction potential, n_c and J, are not evaluated within this framework.

Applied to large interacting systems, our non-Markovian maximum-likelihood method performs well even for relatively short trajectories. Taking, for instance, trajectories of length $L = 200$ for systems of $N = 1000$ particles already enables us to achieve good accuracy, with undistinguishable features in the inference of η/χ and T/χ compared to the non-Bayesian method (see Fig. 5). As already pointed out, the need for very long trajectories in the max-likelihood scheme, for both single-particle and many-particle models, stems from two different facts. First, the shift-invariance approximation introduced by enforcing a Toeplitz structure for the noise covariance matrix results in better performance for longer trajectories. Second, the empirical reconstruction of two-time correlations, which are the quantities that enter into inference formulas, improves when achieved from longer trajectories as compared to shorter ones. In other words, the larger the number of data points, the higher the amount of available information. The advantage of moving from the single oscillator to the many-body interacting case is that a restricted number of "local" quantities turn out to dominate and self-average in sufficiently large systems. So the statistical issue can be at least partially mitigated by averaging over the sample size rather than relying only on temporal averages as we are compelled to do in the case of the harmonic oscillator.

VI. EFFECT OF EXPERIMENTAL ERRORS

So far, we haven't included observation errors in the developed inference scheme, assuming that stochastic trajectories are sampled with infinite accuracy. However, data are typically affected by accuracy limitations and other sources of experimental errors. In the current section, we show the effects of an additional source of noise on the estimation of the model parameters.

The simplest (still realistic, in many practical cases) way to model experimental errors is through a superposition of the discretely sampled trajectory with a sequence of IID Gaussian random variables $\mathcal{N}(0, \sigma^2)$.

As pointed out by several authors [11,12,14], even when σ^2 is very small, measurement noise can impact dynamical inference. A large modification of the high-frequency region of the power spectrum of reconstructed velocities is introduced [14], which, in turn, results in a diverging bias in parameter estimation as $\Delta t \to 0$ [12]. This bias and its trend with Δt appear also in our inference method (see Fig. 6). Intuitively, the inference procedure relies on the increments of the measured degree of freedom, Δx, whose average absolute value has a monotonic dependence on Δt, and need to be compared with the amplitude of measurement errors σ, which we assume to be independent of the data acquisition sampling rate. At very high sampling rates, experimental errors will dominate over the effective dynamics, resulting into an artificial trend $\sim\Delta t^{-1}$ for the parameter η and $\sim\Delta t^{-2}$ for the effective temperature and pulsation of the harmonic oscillator (the same dimensional analysis argument can be extended to the parameters of the inertial spin model).

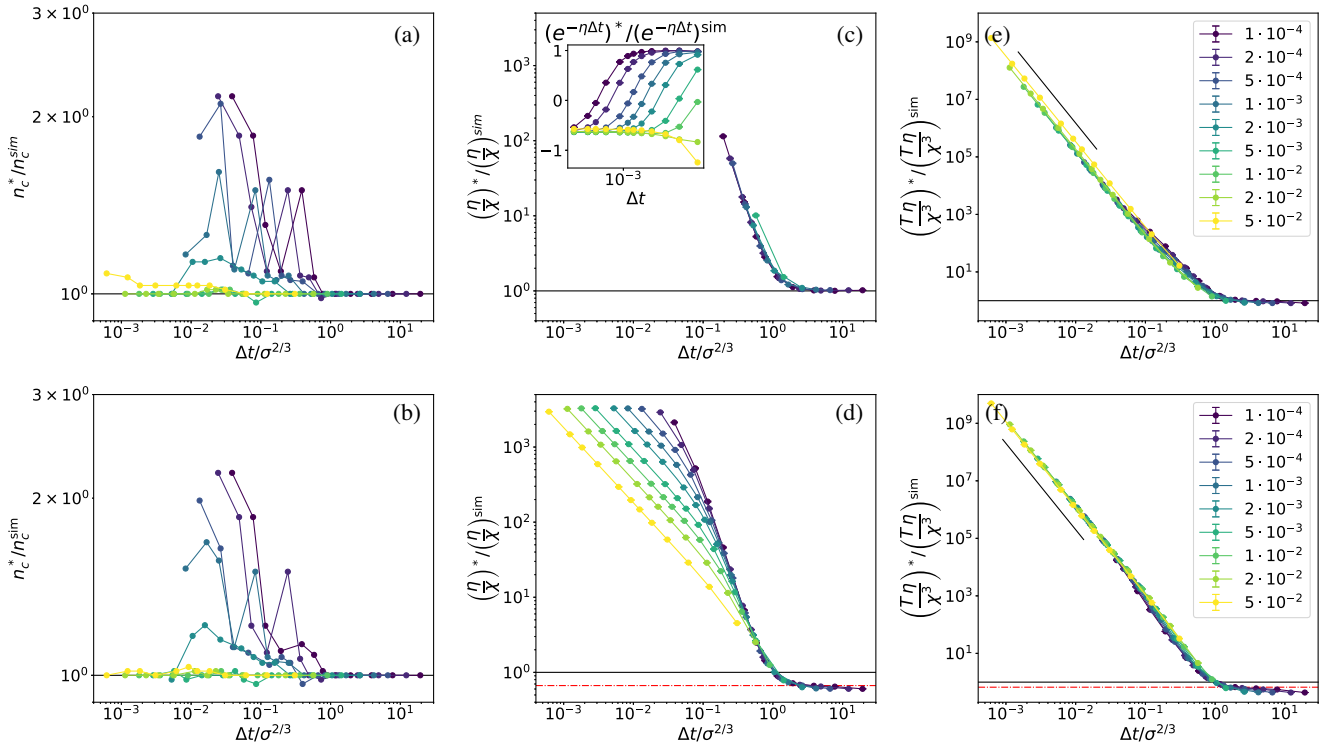

FIG. 6. Effect of measurement error on some selected parameter estimators for the ISM. In the top row, results from the Toeplitz inference scheme are reported; in the bottom row, results from the BBK inference scheme are reported. The rescaling of the inverse sampling rate in the abscissa make the curves in (c)–(f) depart at the same point ($\Delta t/\sigma^{2/3} \sim 1$) from the expected value in the absence of experimental errors (1 for the Toeplitz method; 2/3, marked by the red dot-dashed line, for the BBK method). The collapse of the curves shown in (e) and (f) proves that the control parameter is the ratio between stochastic and experimental noise: $T\eta\Delta t^3/\sigma^2$. The black lines, having a slope of -3, are a guide for the eye. We notice that, for a large noise-to-signal ratio, the estimate of η with the Toeplitz method may be problematic, since estimators of positive definite quantities built with noisy data can become negative, as visible in the inset in (c). We refer to Appendix B 2 for details on the inference formulas. Error bars on (a) and (b) are not shown, for the sake of clarity, whereas in (c)–(f) the 0.95 CI is smaller than the marker size.

Since noise cannot be ignored, we include it in the model in the form of a hidden (non-)Markov model. Suppose we measure noisy discrete data points $(\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_L)$ corresponding to trajectory points $(x_0, x_1, \ldots, x_L)$. Following a maximum-likelihood argument, we estimate the parameters $\boldsymbol{\lambda}$ of the dynamical hidden model as

$$\begin{aligned}\boldsymbol{\lambda}_H^* &= \arg\max_{\boldsymbol{\lambda}} P(\boldsymbol{\lambda}|\hat{x}_0, \ldots, \hat{x}_L)\\ &= \arg\max_{\boldsymbol{\lambda}} P(\hat{x}_0, \ldots, \hat{x}_L|\boldsymbol{\lambda}),\end{aligned} \tag{42}$$

where

$$\begin{aligned}P(\hat{x}_0, \ldots, \hat{x}_L|\boldsymbol{\lambda}) = \int dx_0 \ldots dx_L P(x_0, \ldots, x_L|\boldsymbol{\lambda})\\ \cdot \prod_{n=0}^{L} P(\hat{x}_n|x_n).\end{aligned} \tag{43}$$

We assume $P(\hat{x}_n|x_n) = (1/\sqrt{2\pi\sigma^2}) \exp -((\hat{x}_n - x_n)^2/2\sigma^2)$, and $P(x_0, \ldots, x_L|\boldsymbol{\lambda})$ is determined by the hypothesized dynamical model. As long as we deal with linear models, as in the interacting and noninteracting cases considered above, $P(\hat{x}_0, \ldots, \hat{x}_L|\boldsymbol{\lambda})$ reduces to Gaussian integrals, and the marginalization over the hidden variables can be performed explicitly. A full treatment at any noise-to-signal ratio is then possible but not easily generalizable beyond the harmonic case. For this reason, here we limit ourselves to showing the predicted effect of experimental uncorrelated noise on numerical simulations. Explicit rewriting of the likelihood in Eq. (43) allows us to identify the combination of parameters that control the transition from the small to large noise regime. If $T\eta\Delta t^3/\sigma^2 \ll 1$, noise dominates and, to lowest order, $P(\hat{x}_0, \ldots, \hat{x}_L|\boldsymbol{\lambda}) \simeq \prod_{n=0}^{L}(1/\sqrt{2\pi\sigma^2}) \exp -(\hat{x}_n^2/2\sigma^2)$. If $T\eta\Delta t^3/\sigma^2 \gg 1$, the effect of noise is small, and the likelihood converges to the one we find in the absence of experimental errors.

We conclude, in agreement with Refs. [11,12], that, whenever the experimental apparatus and the observed process are such that the chain of conditions $\sigma^2 \ll T\eta\Delta t^3 \ll 1$ holds, the developed inference strategy still provides a reliable methodology to infer the parameters of

the dynamics. When that condition is not fulfilled, controlled denoising procedures or inference strategies based on hidden modeling must be employed.

VII. CONCLUSIONS

We proposed a maximum-likelihood inference strategy to tackle the problem of learning the best continuous inertial stochastic model from time lapse recordings of an observed process. The problems arising in this context are general, as they stem from the combination of the following three ingredients: the second- (or higher) order nature of the process, when described in terms of the directly measurable degrees of freedom, stochasticity, and the use of discrete sequences of data points. Because of that, contrary to first-order processes, reconstructing the continuous-time dynamical model from the data is not a straightforward task in the case of second-order dynamics. Careful attention must be paid to the mathematical peculiarities of Brownian motion and, in particular, to the minimum order of convergence of the Taylor-Itô expansion allowing for a correct description of infinitesimal fluctuations.

We want a robust inference methodology which could be applied to a wide class of inertial processes, without knowing their exact time-dependent solution. Such a method must then exploit only the local dynamical information carried by the differential equation. Locally in time, the statistical properties of a Markovian or non-Markovian process are determined by the random variable appearing in the discretized Langevin or generalized Langevin equation, respectively. It is then crucial to evaluate correctly the incremental fluctuations, at least to leading order in Δt.

In the considered non-Markovian scenario, the minimum order of convergence required for the Taylor-Itô expansion is $O(\Delta t^{3/2})$. We showed that lower-order approximations lead to the emergence of a $2/3$ rescaling factor for the inferred damping coefficient, as already pointed out in Ref. [26] and in Refs. [12,14] in non-Bayesian settings. Employing known numerical integration schemes [39,40], we developed, to the best of our knowledge, the first max-likelihood inference approach for non-Markovian dynamics (or, equivalently, partially observed Markovian dynamics, since the Markovian embedding is not exploited). We demonstrated the robustness and wide applicability of the method by applying it to different processes: an exactly solvable stochastic oscillator with additive noise with a Gaussian propagator; the Brownian motion of a particle in an anharmonic potential in thermal contact with a heat bath at constant temperature T; a stochastic harmonic oscillator driven by multiplicative noise. While the first two examples are described in equilibrium by Gibbs-Boltzmann distribution, the latter is intrinsically out of equilibrium. In all these cases, our maximum-likelihood estimators for the model parameters are in good agreement with the values used in simulations.

The method can also be successfully and efficiently applied to large interacting systems, with prior modeling of the interaction mechanism. It is in this aspect that the most promising applications of our max-likelihood method possibly lie. The class of processes for which the method has been developed correspond to the simplest way of incorporating memory effects in the equilibrium dynamics of complex Hamiltonian systems. Its fundamental ingredients are linear dissipation and additive noise. With these conditions fulfilled, the problem is computationally efficient and tractable. For noninteracting systems, we showed it is possible to generalize the Bayesian inference approach to nonequilibrium processes driven by multiplicative noise. This generalization should work also for interacting ones.

An important remark is that in this setting only single-valued parameters can be inferred. Heterogeneities in time and space are not taken into account. The proposed method is able to cope with slow time dependence of the parameters compared to the available experimental frame rate, by assuming effectively constant parameters along long subtrajectories. For fast-varying parameters, a better approach is to describe the parameter as a random variable drawn from an unknown distribution and infer the parameters of this distribution. For the spatial heterogeneity in very large systems, unless it is modeled using a small number of parameters, a brute-force maximum-likelihood approach is not feasible and more sophisticated strategies must be developed, as for static inverse problems [16].

Another possible extension of the method is to include a position-dependent dissipation coefficient. This modification would not alter the Gaussian nature of the propagator at short times, even if the noise covariance matrix will no longer be tridiagonal. Nonlinearities in the first derivative of the measured degree of freedom x and v-dependent multiplicative noise could also be considered. Finally, one could try to generalize the approach to higher-order processes, provided that this generalization is motivated by some experimental evidence.

Relating the exact maximum-likelihood procedure to alternative effective inference schemes, like Gloter's minimum contrast strategy [26], also remains an open question. Specifically, is it possible to associate to these non-Markovian processes an effective Markovian description with uncorrelated noise (corresponding to factorized dynamical likelihood) and rescaled parameters? Our analysis suggests that it should be possible to adopt, even if incorrectly, one of the naïve methods discussed in Sec. II B and introduce an *a posteriori* correction of the wrongly estimated parameter, to take into account the effect of the lowest-order discretization.

Another interesting development would be to provide a reliable inference method even in the presence of strong measurement errors. The maximum-likelihood framework

provides a natural formulation for the problem in terms of hidden Markov models.

The natural use of the developed framework is application to real data. Technical specifications of acquisition systems have remarkably improved in the past decades, and it is now possible to collect well-resolved trajectories for long enough time windows. This possibility also holds for animal groups on the move, where experiments are performed in the field and strong limitations are usually set on the acquisition length due to global motion. We know from previous work that the emergent dynamics of groups of birds is dominated by an effective rotational inertia [6]. This inertia allows information to propagate linearly and in an almost undamped way, allowing flocks to turn coherently. Retrieving the effective damping coefficient in this case will allow us to predict the scales where damping becomes relevant, setting a size limit for groups able to collectively change direction. In the context of swarm dynamics, recent theoretical findings [48,49] suggest that the value of the damping coefficient sets—again—a size crossover for groups displaying different critical behavior on the large scale. Understanding the interplay between size, information propagation, and response is a key issue in collective behavior, and a reliable inference approach is crucial to provide well-grounded answers to these questions.

ACKNOWLEDGMENTS

F. F. thanks M. Baldovin for helpful discussions and suggestions. I. G. and F. F. also thank A. Cavagna, A. Vulpiani, and M. Viale. I. G. thanks E. Aurell, S. Bo, and R. Eichhorn. This work was partially supported by the European Research Council Consolidator Grant No. 724208, the European Research Council Advanced Grant No. 785932, and the Italian Ministry of Foreign Affairs and International Cooperation through the Adinmat project.

APPENDIX A: DISCRETIZATION PROCEDURE

Let us briefly summarize two possible systematic strategies to obtain a discretized equation in the space of the x variables up to the desired $O(\Delta t^{3/2})$ order. Following Ref. [8], we can derive from Eq. (3) the associated GLE by formally solving the second equation of the system:

$$v(t) = \int_0^t ds e^{-\eta(t-s)}[f(x(s)) + \xi(s)] + v_0 e^{-\eta t}. \quad \text{(A1)}$$

Plugging this expression back into the equation for x, we get a closed equation in x space:

$$\dot{x} = \int_0^t ds K(t-s) f(x(s)) + \zeta(t) + v_0 e^{-\eta t}, \quad \text{(A2)}$$

where $K(t) = e^{-\eta t}$ and $\zeta(t) = \int_0^t ds e^{-\eta(t-s)}\xi(s)$. Discrete update equations on the scale Δt can now be obtained by integrating Eq. (A2) between t_n and t_{n+1} and between t_{n-1} and t_n. An exponentially decaying memory kernel propagates both the noise and the initial condition v_0 in Eq. (A2); it is then possible to identify an appropriate reweighing of its integrated counterparts in order to get rid of both effects. Indeed, the combination $x_{n+1} - x_n - e^{-\eta\Delta t}(x_n - x_{n-1})$ does not contain v_0 and has a short correlated effective noise:

$$x_{n+1} - x_n - e^{-\eta\Delta t}(x_n - x_{n-1}) = \frac{1 - e^{-\eta\Delta t}}{\eta} \int_{t_{n-1}}^{t_{n+1}} \Psi(t - t_n) f(x(t)) dt + \zeta_n \quad \text{(A3)}$$

with

$$\zeta_n = \int_{t_{n-1}}^{t_{n+1}} \Psi(t - t_n)\xi(t) dt; \quad \Psi(t) = \frac{e^{\eta t} - e^{-\eta\Delta t}}{1 - e^{-\eta\Delta t}}[\theta(t + \Delta t) - \theta(t)] + \frac{1 - e^{\eta(t-\Delta t)}}{1 - e^{-\eta\Delta t}}[\theta(t) - \theta(\Delta t - t)], \quad \text{(A4)}$$

$\theta(t)$ being the Heaviside function. We can check that $\langle \zeta_n \zeta_m \rangle$ has the nearest-neighbor structure of Eq. (26):

$$\langle \zeta_n \zeta_m \rangle = C_{nm} = a\delta_{n,m} + b\delta_{n,m\pm 1}.$$

From Eq. (A4), one deduces that, to the order of $O(\Delta t^3)$, the coefficients a and b of the covariance matrix assume the expression reported in Eq. (27).

So far, these equations are exact. Some approximation is needed at this stage to evaluate the integral of the force. Various methods have been investigated in the literature; among the simplest is the Langevin impulse method [40], which approximates the integral with the function at the midpoint, leading to

$$x_{n+1} = x_n + e^{-\eta\Delta t}(x_n - x_{n-1}) + \frac{1 - e^{-\eta\Delta t}}{\eta} \Delta t f(x_n) + \zeta_n. \quad \text{(A5)}$$

Alternatively, taking the first-order expansion of the force around the midpoint, $f(x(t)) \simeq f(x_n) + (t - t_n) \times [f(x_n) - f(x_{n-1})]/\Delta t$ in Eq. (A3), one recovers the stochastic Verlet algorithm [41], which is one order more accurate than Eq. (A5).

Another approach, followed in Ref. [39] (see also Ref. [50]), is to consider the full system of equations in the (x, v) phase space in integral form:

$$x_{n+1} = x_n + \int_{t_n}^{t_{n+1}} v(t)dt,$$
$$v_{n+1} = v_n + \int_{t_n}^{t_{n+1}} f(x(t))dt + \int_{t_n}^{t_{n+1}} \xi(t)dt \tag{A6}$$

and perform a second-order Taylor-Itô expansion around the point t_n:

$$x_{n+1} = x_n + v_n \Delta t + D_n,$$
$$v_{n+1} = (1 - \eta\Delta t)v_n + \frac{1}{2}\Delta t[f(x_{n+1}) + f(x_n)] + \sigma\Delta t^{1/2}\xi_n - \eta D_n, \tag{A7}$$

where D_n is defined as follows:

$$D_n = \frac{1}{2}\Delta t^2[f(x_n) - \eta v_n] + \sigma\Delta t^{3/2}\left[\frac{1}{2}\xi_n + \frac{1}{2\sqrt{3}}\theta_n\right] \tag{A8}$$

and ξ_n and θ_n are IID Gaussian variables sampled from $\mathcal{N}(0, 1)$. Eliminating the velocity variables v_n and v_{n-1}, we find a difference equation of the form of Eq. (21):

$$x_{n+1} = 2x_n - x_{n-1} - \eta\Delta t(1 - \eta\Delta t/2)(x_n - x_{n-1}) - \Delta t^2 f(x_n) + \frac{\eta\Delta t^3}{4}[f(x_n) + 3f(x_{n-1})] + \Delta t^{3/2}\zeta_n \tag{A9}$$

with α and β coinciding, up to $O(\Delta t^2)$, with the Taylor expansion of the coefficients in Eq. (A5). The noise variable ζ_n is defined from Eq. (A7) as a linear combination of ξ_n, ξ_{n-1}, θ_n, and θ_{n-1}. As a result, due to overlapping Wiener processes, correlations between subsequent noise extractions emerge, which are still described by Eq. (26).

This second derivation is helpful in justifying the quasi-Toeplitz structure of the covariance matrix discussed in the main text. Indeed, fixing x_1 implies fixing the first random increment which is responsible for position update in the integration scheme Eq. (A7), when the known initial conditions are (x_0, v_0). Since this stochastic increment enters into the definition of ζ_1 but not in that of ζ_2, the true covariance matrix must have a different entry C_{11} than the other elements on the main diagonal, as in Eq. (28).

APPENDIX B: INFERENCE FORMULAS

1. Naïve max-likelihood approaches for the harmonic oscillator

Several Euler-like schemes for the numerical integration of second-order stochastic differential equations can be defined. From each of them, inconsistently retaining only the diagonal $O(\Delta t^{3/2})$ stochastic terms when we write the update equations in x space, we can extract a factorized expression for the dynamical likelihood, such as Eq. (10).

Let us focus on three particular examples (shown in Fig. 7 for the harmonic oscillator and Fig. 8 for the ISM): the standard explicit Euler-Maruyama scheme (EM-fwd), its implicit variant (EM-bkd), and the symmetric Brünger-Brooks-Karplus (BBK) scheme [51]. The three of them may be obtained from the second-order SDE Eq. (1) by approximating first and second time derivatives adopting a forward, backward, or symmetric prescription, respectively. The resulting update equations in the three cases read

$$\text{[EM-fwd]} \quad x_{n+1} - (2 - \eta\Delta t)x_n + (1 - \eta\Delta t + \omega_0^2\Delta t^2)x_{n-1} = \sigma\Delta t^{3/2} r_{n-1}, \tag{B1}$$

$$\text{[EM-bkd]} \quad (1 + \eta\Delta t)x_{n+1} - (2 + \eta\Delta t - \omega_0^2\Delta t^2)x_n + x_{n-1} = \sigma\Delta t^{3/2} r_{n+1}, \tag{B2}$$

$$\text{[BBK]} \quad \left(1 + \frac{\eta\Delta t}{2}\right)x_{n+1} - (2 - \omega_0^2\Delta t^2)x_n + \left(1 - \frac{\eta\Delta t}{2}\right)x_{n-1} = \sigma\Delta t^{3/2} r_n \tag{B3}$$

with $\sigma = \sqrt{2T\eta}$ and $\{r_n\}$ a sequence of $L - 1$ IID Gaussian random variables of null mean and unit variance.

Thanks to the independence of the random variables appearing in Eqs. (B1)–(B3), the discrete propagator takes an approximate factorized form, which we can generally write as

$$P_{(1)}(x_2, \ldots, x_L | x_0, x_1) = \prod_{n=1}^{L-1} \frac{1}{Z_n} e^{-S_n(x_{n+1}, x_n, x_{n-1})}. \tag{B4}$$

The reduced minus-log-likelihood, defined as

$$\frac{\mathcal{L}}{L-1} := \frac{-\ln P(x_2, \ldots, x_L | x_0, x_1)}{L-1}, \tag{B5}$$

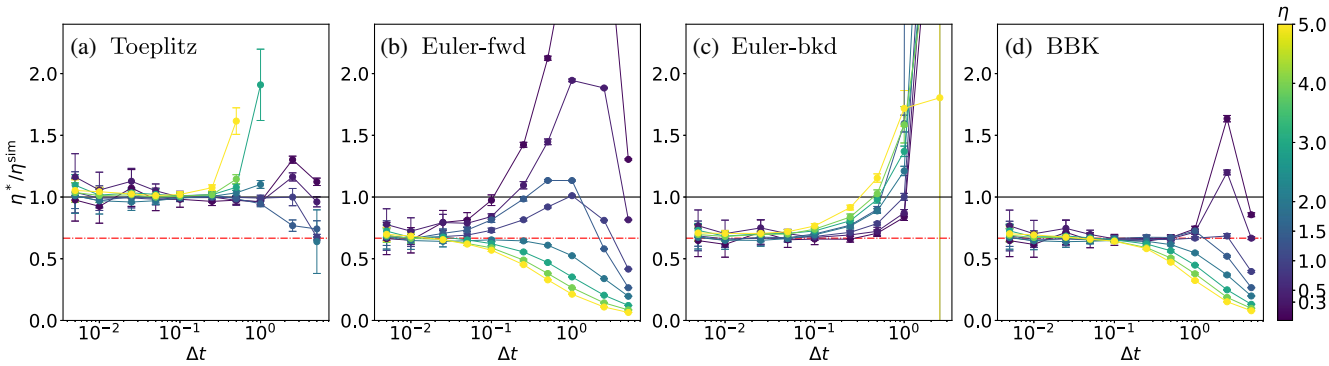

FIG. 7. Accuracy of the different likelihood-based methods in inferring the damping coefficient of the harmonic oscillator, in varying damping regimes: (a) shows the $O(\Delta t^{3/2})$ Toeplitz method; (b)–(d) show the three $O(\Delta t^{1/2})$ variants corresponding, respectively, to the Euler-forward, Euler-backward, and BBK schemes. The main features to highlight are the appearance of the $2/3$ rescaling factor for the $O(\Delta t^{1/2})$ scheme (red dot-dashed line), compared to the absence of any rescaling for the $O(\Delta t^{3/2})$ scheme, and the higher stability of the latter with respect to Δt filtering. Different damping regimes are explored: The sampled values of $\eta^{\rm sim}$ are indicated in the color bar. The remaining parameters are $T = 1$ and $\omega_0 = 1$. Each point is the average of the inference results of ten different trajectories of 5000 points (for any Δt). Error bars are taken as 0.95 CI.

corresponds in the factorized case to the temporal average of the quantity $S_n + \ln Z_n$. This quantity is defined in a slightly different way in the three cases above; consequently, in each of these cases, the reduced minus-log-likelihood is slightly different, as reads in the following. We recall the notation used in the main text to indicate the equal-time, one-step, and two-step experimental correlation functions:

$$C_s = \frac{1}{L-1}\sum_{n=1}^{L-1} x_n x_n; \qquad C_s' = \frac{1}{L-1}\sum_{n=1}^{L-1} x_{n+1} x_{n+1};$$

$$C_s'' = \frac{1}{L-1}\sum_{n=1}^{L-1} x_{n-1} x_{n-1}; \qquad G_s = \frac{1}{L-1}\sum_{n=1}^{L-1} x_n x_{n+1};$$

$$G_s' = \frac{1}{L-1}\sum_{n=1}^{L-1} x_n x_{n-1}; \qquad F_s = \frac{1}{L-1}\sum_{n=1}^{L-1} x_{n-1} x_{n+1}.$$

$$[\text{EM-fwd}]\,\frac{\mathcal{L}}{L-1} = \frac{1}{2}\ln\left(4\pi T\eta\Delta t^3\right) + \frac{1}{4T\eta\Delta t^3}\left[C_s' + (2-\eta\Delta t^2)^2 C_s + (1-\eta\Delta t+\omega_0^2\Delta t^2)^2 C_s'' - 2(2-\eta\Delta t)G_s \right.$$
$$\left. + 2(1-\eta\Delta t+\omega_0^2\Delta t^2)F_s - 2(2-\eta\Delta t)(1-\eta\Delta t+\omega_0^2\Delta t^2)G_s'\right]; \tag{B6}$$

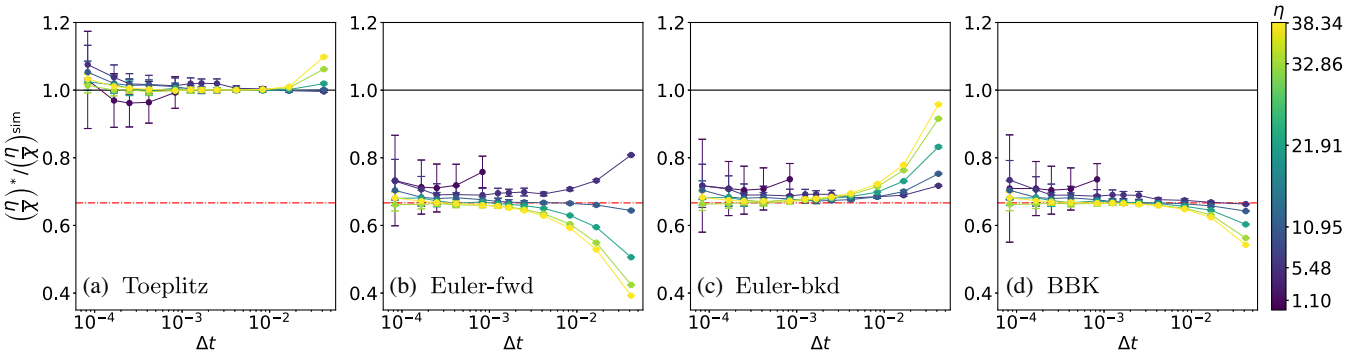

FIG. 8. Accuracy of the different likelihood-based methods in inferring the effective parameter η/χ of the inertial spin model: (a) shows the $O(\Delta t^{3/2})$ Toeplitz method; (b)–(d) show the three $O(\Delta t^{1/2})$ variants corresponding, respectively, to the Euler-forward, Euler-backward, and BBK schemes. We see the $2/3$ factor for the $O(\Delta t^{1/2})$ schemes (red dot-dashed line) and no rescaling for the $O(\Delta t^{3/2})$ scheme. ISM simulations are performed in different damping regimes: The sampled values for the parameter $\eta^{\rm sim}$ are indicated along the color bar. The remaining parameters are $\chi = 1$, $T = 0.4$, $J = 5.0$, $n_c = 6$, and $N = 1000$. Each point is the average of the inference results of ten different trajectories of 200 points (for any Δt). Error bars correspond to 0.95 CI.

$$[\text{EM-bkd}]\frac{\mathcal{L}}{L-1} = \frac{1}{2}\ln(4\pi T\eta\Delta t^3) - \ln(1+\eta\Delta t) + \frac{1}{4T\eta\Delta t^3}[(1+\eta\Delta t)^2 C_s' + (2+\eta\Delta t-\omega_0^2\Delta t^2)^2 C_s + C_s'' - 2(1+\eta\Delta t)(2+\eta\Delta t-\omega_0^2\Delta t^2)G_s + 2(1+\eta\Delta t)F_s - 2(2+\eta\Delta t-\omega_0^2\Delta t^2)G_s'];\tag{B7}$$

$$[\text{BBK}]\frac{\mathcal{L}}{L-1} = \frac{1}{2}\ln(4\pi T\eta\Delta t^3) - \ln\left(1+\frac{\eta\Delta t}{2}\right) + \frac{1}{4T\eta\Delta t^3}[(1+\eta\Delta t/2)^2 C_s' + (2-\omega_0^2\Delta t^2)^2 C_s + (1-\eta\Delta t/2)^2 C_s'' - 2(2-\omega_0^2\Delta t^2)(1+\eta\Delta t/2)G_s + 2(1+\eta\Delta t/2)(1-\eta\Delta t/2)F_s - 2(2-\omega_0^2\Delta t^2)(1-\eta\Delta t/2)G_s'].\tag{B8}$$

Minimization of Eqs. (B6)–(B8) with respect to the parameters of the model yields the following optimal values, according to the adopted scheme.

(i) Euler-forward:

$$\eta^*_{\text{fwd}} = \frac{1}{\Delta t}\frac{G_s + G_s' - 2C_s + \frac{G_s'}{C_s''}(2G_s' - C_s'' - F_s)}{-C_s + \frac{G_s'^2}{C_s''}};\tag{B9}$$

$$\omega^{2*}_{0\,\text{fwd}} = \frac{1}{\Delta t^2}\frac{(2-\eta\Delta t)G_s' - (1-\eta\Delta t)C_s'' - F_s}{C_s''};\tag{B10}$$

$$T^*_{\text{fwd}} = \frac{1}{2\eta\Delta t^3}[C_s' + (2-\eta\Delta t^2)^2 C_s + (1-\eta\Delta t+\omega_0^2\Delta t^2)^2 C_s'' - 2(2-\eta\Delta t)G_s + 2(1-\eta\Delta t+\omega_0^2\Delta t^2)F_s - 2(2-\eta\Delta t)(1-\eta\Delta t+\omega_0^2\Delta t^2)G_s'];\tag{B11}$$

(ii) Euler-backward:

$$\eta^*_{\text{bkd}} = \frac{1}{\Delta t}\frac{C_s'' + F_s - \frac{G_s'}{C_s}(G_s + G_s')}{\frac{G_s G_s'}{C_s} - F_s};\tag{B12}$$

$$\omega^{2*}_{0\,\text{bkd}} = \frac{1}{\Delta t^2}\frac{(2+\eta\Delta t)C_s - G_s' - (1+\eta\Delta t)G_s}{C_s};\tag{B13}$$

$$T^*_{\text{bkd}} = \frac{1}{2\eta\Delta t^3}[(1+\eta\Delta t)^2 C_s' + (2+\eta\Delta t-\omega_0^2\Delta t^2)^2 C_s + C_s'' - 2(1+\eta\Delta t)(2+\eta\Delta t-\omega_0^2\Delta t^2)G_s + 2(1+\eta\Delta t)F_s - 2(2+\eta\Delta t-\omega_0^2\Delta t^2)G_s'];\tag{B14}$$

(iii) BBK:

$$\eta^*_{\text{BBK}} = \frac{2}{\Delta t}\frac{C_s'' + F_s - \frac{G_s'}{C_s}(G_s + G_s')}{C_s'' - F_s - \frac{G_s'}{C_s}(G_s' - G_s)};\tag{B15}$$

$$\omega^{2*}_{0\,\text{BBK}} = \frac{1}{\Delta t^2}\frac{2C_s - (1+\frac{\eta\Delta t}{2})G_s - (1-\frac{\eta\Delta t}{2})G_s'}{C_s};\tag{B16}$$

$$T^*_{\text{BBK}} = \frac{1}{2\eta\Delta t^3}[(1+\eta\Delta t/2)^2 C_s' + (2-\omega_0^2\Delta t^2)^2 C_s + (1-\eta\Delta t/2)^2 C_s'' - 2(2-\omega_0^2\Delta t^2)(1+\eta\Delta t/2)G_s + 2(1+\eta\Delta t/2)(1-\eta\Delta t/2)F_s - 2(2-\omega_0^2\Delta t^2)(1-\eta\Delta t/2)G_s'].\tag{B17}$$

All of the schemes above are derived from numerical integrators with weak and strong convergence order $O(\Delta t)$ and, consequently, give a 2/3 rescaling factor for the inferred damping coefficient η^*. This result can be checked using the procedure outlined to derive Eq. (15), which consists of replacing the experimental two-time quantities with the known correlation functions for the harmonic oscillator:

$$C(t) = \frac{T}{\omega_0^2}e^{-\gamma t}\left[\cos\left(\sqrt{\omega_0^2-\gamma^2}t\right) + \gamma\frac{\sin(\sqrt{\omega_0^2-\gamma^2}t)}{\sqrt{\omega_0^2-\gamma^2}}\right],\tag{B18}$$

where $\gamma = \eta/2$, and performing a Taylor expansion around the zero temporal distance. In the same way, the exactness of the inference formulas for T^* and ω_0^{2*} can be checked for the three methods.

2. Shift-invariant $O(\Delta t^{3/2})$ Bayesian approach

We argue that the joint probability of sequences of points in real space is not factorized into a chain of

conditional probabilities. Indeed, the dynamics of the harmonic oscillator, when projected into the x space, is governed by evolution equations containing a colored noise. The right scheme to adopt is then of the kind of Eq. (21): As discussed in the main text, this scheme requires correlations between subsequently extracted random variables to be taken into account, resulting, in the case of additive noise, in a covariance matrix with a (quasi-)Toeplitz symmetric tridiagonal structure [cf. Eqs. (28) and (26)]. We pursue a maximum-likelihood approach, taking as the function of the parameters of the model to maximize

$$P_{(2)}(x_L,\ldots,x_2|x_1,x_0)=\frac{1}{Z}\exp-\frac{1}{2}\sum_{n,m=1}^{L-1}[x_{n+1}+F(x_n,x_{n-1};\boldsymbol{\mu})]C^{-1}{}_{nm}[x_{m+1}+F(x_m,x_{m-1};\boldsymbol{\mu})]. \tag{B19}$$

The partition function is specified by Eqs. (24) and (30), whereas the relation between $\boldsymbol{\mu}$ and the physical parameters of the dynamical model depends on the details of the discretization scheme which is adopted.

Thanks to the peculiar structure of this likelihood, one can go pretty far with simple algebra in the optimization problem. First of all, it is convenient to reformulate the issue as a minimization problem for the minus-log-likelihood:

$$L=\frac{L-1}{2}\ln\left(2\pi\frac{2}{3}T\eta\Delta t^3\right)+\frac{1}{2}\sum_{k=1}^{L-1}\ln\left[2+\cos(k\pi/L)\right]+\frac{3/2}{LT\eta\Delta t^3}\sum_{n,m=1}^{L-1}[x_{n+1}+F(x_n,x_{n-1};\mathbf{\mu})]\tilde{A}_{nm}[x_{m+1}+F(x_m,x_{m-1};\mathbf{\mu})], \tag{B20}$$

being

$$\tilde{A}_{nm}=\sum_{k=1}^{L-1}\frac{\sin(\frac{nk\pi}{L})\sin(\frac{mk\pi}{L})}{2+\cos(\frac{k\pi}{L})}. \tag{B21}$$

As usual, the temperature just appears as a prefactor for the effective action, without affecting its actual dynamical structure. The optimal value is given by

$$T^*=\frac{3}{L(L-1)\eta\Delta t^3}\sum_{n,m=1}^{L-1}[x_{n+1}+F(x_n,x_{n-1};\mathbf{\mu})]\tilde{A}_{nm}[x_{m+1}+F(x_m,x_{m-1};\mathbf{\mu})]. \tag{B22}$$

Replacing it into Eq. (B20) and getting rid of additional constants, we obtain a reduced minus-log-likelihood:

$$\mathcal{L}\propto\frac{1}{L-1}\sum_{n,m=1}^{L-1}\sum_{k=1}^{L-1}\frac{\sin(\frac{nk\pi}{L})\sin(\frac{mk\pi}{L})}{2+\cos(\frac{k\pi}{L})}[x_{n+1}+F(x_n,x_{n-1};\boldsymbol{\mu})][x_{m+1}+F(x_m,x_{m-1};\boldsymbol{\mu})]. \tag{B23}$$

One can now split all the terms appearing in the sum and derive with respect to the effective parameters $\boldsymbol{\mu}$. Focusing on the case of the simple stochastic harmonic oscillator, $F(x_n,x_{n-1};\boldsymbol{\mu})=\alpha x_n+\beta x_{n-1}$, we have $\boldsymbol{\mu}=(\alpha,\beta)$. By adopting the Langevin impulse integrator (see Appendix A), the effective parameters correspond to

$$\alpha=-1-e^{-\eta\Delta t}+\omega_0^2\Delta t(1-e^{-\eta\Delta t})/\eta,\qquad \beta=e^{-\eta\Delta t}. \tag{B24}$$

By adopting a second-order Taylor expansion around the prepoint, they correspond to

$$\alpha=-2+\eta\Delta t\left(1-\frac{\eta\Delta t}{2}\right)+\omega_0^2\Delta t^2,\qquad \beta=1-\eta\Delta t\left(1-\frac{\eta\Delta t}{2}\right). \tag{B25}$$

As required for them to be consistent, the two variants are equivalent up to $O(\Delta t^3)$. The numerical results shown in this paper are obtained using Eq. (B24).

Imposing that the derivatives of $\mathcal{L}$ with respect to α and β are zero leads to

$$\alpha^*=-\frac{T_1+\beta^*T_3}{2T_4};\qquad \beta^*=\frac{T_1T_3-2T_2T_4}{-T_3^2+4T_4T_5}, \tag{B26}$$

where, with an implicit sum over the indexes n, m from 1 to $L-1$,

$$T_1 = \frac{2}{L}\tilde{A}_{nm}x_n x_{m+1}; \qquad T_2 = \frac{2}{L}\tilde{A}_{nm}x_{n-1}x_{m+1};$$
$$T_3 = \frac{2}{L}\tilde{A}_{nm}x_{n-1}x_m; \qquad T_4 = \frac{1}{L}\tilde{A}_{nm}x_n x_m$$
$$T_5 = \frac{1}{L}\tilde{A}_{nm}x_{n-1}x_{m-1}. \tag{B27}$$

This procedure can be applied to find explicit formulas for any noninteracting system described by a Kramers process with velocity-independent forces $f(x)$, as in Eq. (1). We report here those we derive and use for the anharmonic model with force $f(x) = -kx - \lambda x^3$. Referring again to the Langevin impulse integrator, one possible set of parameters is given by $\boldsymbol{\mu} = (\beta, K, \Lambda)$, where $\beta = e^{-\eta\Delta t}$, $K = k\Delta t/\eta$, and $\Lambda = \lambda\Delta t/\eta$. The Toeplitz inference formulas for those parameters read

$$\beta^* = \left[P_5 - \frac{P_6P_8}{P_2} - \frac{(P_2P_9 - P_3P_8)(P_2P_7 - P_3P_6)}{P_2(P_2P_4 - P_3^2)}\right] \cdot \left[P_1 - \frac{P_6^2}{P_2} - \frac{(P_2P_7 - P_3P_6)^2}{P_2(P_2P_4 - P_3^2)}\right]^{-1}; \tag{B28}$$

$$\Lambda^* = \frac{\beta^*(P_2P_7 - P_3P_6) - (P_2P_9 - P_3P_8)}{(1-\beta^*)(P_2P_4 - P_3^2)}; \tag{B29}$$

$$K^* = \frac{\beta^* P_6 - P_8}{(1-\beta^*)P_2} - \Lambda^*\frac{P_3}{P_2}; \tag{B30}$$

where

$$P_1 = (x_n - x_{n-1})\tilde{A}_{nm}(x_m - x_{m-1});$$
$$P_2 = x_n\tilde{A}_{nm}x_m; \qquad P_3 = x_n\tilde{A}_{nm}x_m^3; \qquad P_4 = x_n^3\tilde{A}_{nm}x_m^3;$$
$$P_5 = (x_n - x_{n-1})\tilde{A}_{nm}(x_{m+1} - x_m);$$
$$P_6 = (x_n - x_{n-1})\tilde{A}_{nm}x_m; \qquad P_7 = (x_n - x_{n-1})\tilde{A}_{nm}x_m^3;$$
$$P_8 = (x_{n+1} - x_n)\tilde{A}_{nm}x_m; \qquad P_9 = (x_{n+1} - x_n)\tilde{A}_{nm}x_m^3. \tag{B31}$$

From these equations, the max-likelihood estimators for the physical parameters λ^*, k^*, and η^* can be found.

3. Generalization to the interacting case (ISM)

As one moves from single- to many-particle systems, extra parameters are needed: Position and velocity variables are conveniently represented as Nd-component vectors, N being the number of constituents of the group, while model parameters become matrices. For the equations of motion of the three-dimensional ISM on a fixed lattice in the spin-wave approximation Eq. (41), the update rule becomes

$$\boldsymbol{\zeta}_n^i = \boldsymbol{\pi}_{n+1}^i + \alpha^{ij}\boldsymbol{\pi}_{nj} + \beta^{ij}\boldsymbol{\pi}_{n-1j}, \tag{B32}$$

with

$$\alpha^{ij} = \alpha^0\delta^{ij} + \alpha^1\Lambda^{ij} \quad \text{and} \quad \beta^{ij} = \beta\delta^{ij},$$

where Λ^{ij} is the discrete Laplacian and sums over the j index are implicit.

The definitions of α^0, α^1, and β depend on the details of the discretization. Using, for instance, the Langevin impulse integrator,

$$\left(\frac{\eta}{\chi}\right)^*_{\mathrm{LI}} = -\frac{\ln\beta^*}{\Delta t}; \qquad \left(\frac{J}{\chi}\right)^*_{\mathrm{LI}} = -\frac{\ln\beta^*}{1-\beta^*}\frac{\alpha^*}{\Delta t^2}. \tag{B33}$$

Using instead a second-order Taylor expansion, we get

$$\left(\frac{\eta}{\chi}\right)^*_{\mathrm{IIT}} = \frac{1-\sqrt{2\beta^*-1}}{\Delta t}; \qquad \left(\frac{J}{\chi}\right)^*_{\mathrm{IIT}} = \frac{\alpha^*}{\Delta t^2}. \tag{B34}$$

The three parameters α^0, α^1, and β are not independent, since the extra independent parameters of the interacting problem are hidden in the adjacency matrix. In both of the cases considered above [Eqs. (B33) and (B34)], α^0 and β are linked by the same relation: $\alpha^0 = -\beta - 1$. Renaming $\alpha^1 = \alpha$, the minus-log-likelihood reads

$$\mathcal{L} = \frac{(L-1)(d-1)}{2}\ln\left(\frac{T\eta}{\chi^2}\Delta t^3\right) + \mathrm{const} + \frac{3/2}{L\frac{T\eta}{\chi^2}\Delta t^3}\sum_{n,m=1}^{L-1}\frac{1}{N}\sum_{i=1}^{N}\tilde{A}_{nm}[\boldsymbol{\pi}_{n+1}^i - \boldsymbol{\pi}_n^i - \beta(\boldsymbol{\pi}_n^i - \boldsymbol{\pi}_{n-1}^i) + \alpha\Lambda^{ij}\boldsymbol{\pi}_{nj}] \cdot [\boldsymbol{\pi}_{m+1}^i - \boldsymbol{\pi}_i^m - \beta(\boldsymbol{\pi}_m^i - \boldsymbol{\pi}_{m-1}^i) + \alpha\Lambda^{il}\boldsymbol{\pi}_{ml}]. \tag{B35}$$

Again, one can proceed with an analytic minimization with respect to T, α, and β, giving

$$\alpha^* = \frac{\beta K_4 - K_3}{2K_5}; \tag{B36}$$

$$\beta^* = \frac{-K_3K_4 + 2K_1K_5}{-K_4^2 + 4K_2K_5}; \tag{B37}$$

$$T^* = \frac{3}{(d-1)(\eta/\chi)^*\Delta t^3}[K_0 - \beta^* K_1 + \beta^{*2}K_2 + \alpha^* K_3 - \alpha^*\beta^* K_4 + \alpha^{*2}K_5], \tag{B38}$$

with $K_0 \ldots K_5$ the generalization to the many-particle case of the combinations of experimental observables $T_1, \ldots, T_5$ defined above (again with implicit sums over n, m):

$$K_0 = \frac{1}{N}\sum_{i=1}^{N}\frac{\tilde{A}_{nm}}{L(L-1)}(\boldsymbol{\pi}^i_{n+1} - \boldsymbol{\pi}^i_n)\cdot(\boldsymbol{\pi}^i_{m+1} - \boldsymbol{\pi}^i_m);$$

$$K_1 = \frac{2}{N}\sum_{i=1}^{N}\frac{\tilde{A}_{nm}}{L(L-1)}(\boldsymbol{\pi}^i_{n+1} - \boldsymbol{\pi}^i_n)\cdot(\boldsymbol{\pi}^i_m - \boldsymbol{\pi}^i_{m-1});$$

$$K_2 = \frac{1}{N}\sum_{i=1}^{N}\frac{\tilde{A}_{nm}}{L(L-1)}(\boldsymbol{\pi}^i_n - \boldsymbol{\pi}^i_{n-1})\cdot(\boldsymbol{\pi}^i_m - \boldsymbol{\pi}^i_{m-1});$$

$$K_3 = \frac{2}{N}\sum_{i,j=1}^{N}\frac{\tilde{A}_{nm}}{L(L-1)}\Lambda^{ij}(\boldsymbol{\pi}^i_{n+1} - \boldsymbol{\pi}^i_n)\cdot\boldsymbol{\pi}^j_m;$$

$$K_4 = \frac{2}{N}\sum_{i,j=1}^{N}\frac{\tilde{A}_{nm}}{L(L-1)}\Lambda^{ij}(\boldsymbol{\pi}^i_n - \boldsymbol{\pi}^i_{n-1})\cdot\boldsymbol{\pi}^j_m;$$

$$K_5 = \frac{1}{N}\sum_{i=1}^{N}\frac{\tilde{A}_{nm}}{L(L-1)}\Lambda^{ij}\Lambda^{il}\boldsymbol{\pi}^j_n\cdot\boldsymbol{\pi}^l_m.$$

Results obtained from these formulas are reported in Fig. 8 and in the main text.

4. Generalization to the multiplicative case

From Eqs. (33) and (34), which define the discrete update rule for the multiplicative process described by Eq. (32), one can derive max-likelihood estimators for the parameters of the model by minimizing the effective cost function in Eq. (36). Let us carry on this explicitly for the following reference example:

$$\ddot{x} = -\eta v - kx + \sqrt{a + bx^2}\xi, \tag{B39}$$

such that the quantities appearing in Eq. (36) read

$$F(x_n, x_{n-1}; \boldsymbol{\mu}) = -x_n - e^{-\eta\Delta t}(x_n - x_{n-1}) + (1 - e^{-\eta\Delta t})\frac{k\Delta t}{\eta}x_n, \tag{B40}$$

with $\boldsymbol{\mu} = (e^{-\eta\Delta t}, k\Delta t/\eta)$ and

$$C_{nm} = \frac{2}{3}\Delta t^3(a + bx_n^2)\delta_{n,m} + \frac{1}{6}\Delta t^3\sqrt{(a + bx_n^2)(a + bx_m^2)}\delta_{n,m\pm1}. \tag{B41}$$

Simple manipulations allow us to reduce to the minimization problem to a one-dimensional numerical optimization, since analytical formulas for the optimal values of the effective parameters $\beta = e^{-\eta\Delta t}$, $K = k\Delta t/\eta$, and b can easily be found:

$$b^* = \frac{3/\Delta t^3}{L-1}\left[P_0^* - \frac{P_4^{*2}}{P_2^*} - \frac{(P_3^*P_2^* - P_4^*P_5^*)^2}{P_1^*P_2^* - P_5^{*2}}\right];$$
$$\beta^* = \frac{P_3^*P_2^* - P_4^*P_5^*}{P_1^*P_2^* - (P_5^*)^2}; K^* = \frac{\beta^* P_5^* - P_4^*}{(1-\beta^*)P_2^*}, \tag{B42}$$

where we rename $\alpha = a/b$ and $P_i^* = P_i(\alpha^*)$. The optimal value of the new effective parameter α^* is the minimizer of the following function of α:

$$\mathcal{L} = \frac{1}{L-1}\sum_{k=1}^{L-1}\ln\tilde{\lambda}_k + \ln\left[P_0 - \frac{P_4^2}{P_2} - \frac{(P_3P_2 - P_4P_5)(P_3P_2 - P_4P_5)}{P_1P_2 - P_5^2}\right], \tag{B43}$$

where $\{\tilde{\lambda}_k\}$ is the set of eigenvalues of the reduced covariance matrix

$$A^{-1}{}_{nm} = 3C_{nm}/(b\Delta t^3) = 2(\alpha + x_n^2)\delta_{n,m} + \frac{1}{2}\sqrt{(\alpha + x_n^2)(\alpha + x_m^2)}\delta_{n,m\pm1}, \tag{B44}$$

and

$$P_0 = \frac{1}{L-1}\sum_{n,m=1}^{L-1}(x_{n+1} - x_n)A_{nm}(x_{m+1} - x_m);$$

$$P_1 = \frac{1}{L-1}\sum_{n,m=1}^{L-1}(x_n - x_{n-1})A_{nm}(x_m - x_{m-1});$$

$$P_2 = \frac{1}{L-1}\sum_{n,m=1}^{L-1}x_nA_{nm}x_m;$$

$$P_3 = \frac{1}{L-1}\sum_{n,m=1}^{L-1}(x_{n+1} - x_n)A_{nm}(x_m - x_{m-1});$$

$$P_4 = \frac{1}{L-1}\sum_{n,m=1}^{L-1}(x_{n+1} - x_n)A_{nm}x_m;$$

$$P_5 = \frac{1}{L-1}\sum_{n,m=1}^{L-1}(x_n - x_{n-1})A_{nm}x_m.$$

5. Non-Bayesian approach: Inference formulas without a likelihood

We build in this section an alternative approach to the Bayesian one, as outlined in Sec. III of the main text. To be explicit, we need to choose a discrete update equation in x space: Let us choose again the one corresponding to the usual continuation rule of the LI:

$$x_{n+1} = x_n + e^{-\eta\Delta t}(x_n - x_{n-1}) + \frac{1-e^{-\eta\Delta t}}{\eta}\omega_0^2\Delta t x_n + \zeta_n, \tag{B45}$$

multiply its rhs and lhs by x_n, x_{n+1}, and x_{n-1}, and take the average over the noise distribution. The resulting equations are

$$\langle x_{n+1}x_n\rangle = \langle {x_n}^2\rangle + e^{-\eta\Delta t}(\langle {x_n}^2\rangle - \langle x_n x_{n-1}\rangle) + \frac{1-e^{-\eta\Delta t}}{\eta}\omega_0^2\Delta t\langle {x_n}^2\rangle + \langle x_n\zeta_n\rangle; \tag{B46}$$

$$\langle x_{n+1}x_{n+1}\rangle = \langle x_n x_{n+1}\rangle + e^{-\eta\Delta t}(\langle x_n x_{n+1}\rangle - \langle x_{n-1}x_{n+1}\rangle) + \frac{1-e^{-\eta\Delta t}}{\eta}\omega_0^2\Delta t\langle x_n x_{n+1}\rangle + \langle\zeta_n x_{n+1}\rangle; \tag{B47}$$

$$\langle x_{n+1}x_{n-1}\rangle = \langle x_n x_{n-1}\rangle + e^{-\eta\Delta t}(\langle x_n x_{n-1}\rangle - \langle x_{n-1}^2\rangle) + \frac{1-e^{-\eta\Delta t}}{\eta}\omega_0^2\Delta t\langle x_n x_{n-1}\rangle. \tag{B48}$$

Using again Eq. (B45)—combined with the covariance matrix of the Gaussian variables—to compute $\langle\zeta_n x_n\rangle$ and $\langle\zeta_n x_{n+1}\rangle$, the relations we find are

$$G_s = C_s + e^{-\eta\Delta t}(C_s - G_s') + \frac{1-e^{-\eta\Delta t}}{\eta}\omega_0^2\Delta t C_s + b; \tag{B49}$$

$$C_s' = G_s + b + a + e^{-\eta\Delta t}(G_s - F_s + b) + \frac{1-e^{-\eta\Delta t}}{\eta}\omega_0^2\Delta t(G_s + b); \tag{B50}$$

$$F_s = G_s' + e^{-\eta\Delta t}(G_s' - C_s'') + \frac{1-e^{-\eta\Delta t}}{\eta}\omega_0^2\Delta t G_s'. \tag{B51}$$

In order to find Eqs. (B49)–(B51), we identify the actual correlation functions with the empirical ones, denoted with C, G, and F symbols, and we hypothesize a stationarity assumption to hold to explicitly compute them. After proper manipulation, one can extract "inference relations" for b, $e^{-\eta\Delta t}$, and $\omega_0^2\Delta t$ and derive from them the physical parameters of the model. In order, $e^{-\eta\Delta t}$ is given as the solution of the second-degree polynomial equation:

$$\begin{aligned}&(2G_s' - C_s - C_s'')e^{-2\eta\Delta t} + [2G_s + C_s'' - C_s\\ &\quad -2F_s + 5(2G_s' - C_s - C_s'')]e^{-\eta\Delta t} + [G_s\\ &\quad -G_s' + F_s - C_s' + 5(G_s' - C_s - F_s + G_s)] = 0;\end{aligned} \tag{B52}$$

then b and $\omega_0^2\Delta t$ are computed as follows:

$$b = G_s' - F_s + G_s - C_s + e^{-\eta\Delta t}(2G_s' - C_s'' - C_s); \tag{B53}$$

$$\omega_0^2\Delta t = \frac{-\eta}{1-e^{-\eta\Delta t}}\left[\frac{G_s - C_s - b}{C_s} - e^{-\eta\Delta t}\frac{C_s - G_s'}{C_s}\right]. \tag{B54}$$

Notice that these inference equations are not unique. Combining the starting equations in a different way results in slightly different inference formulas, which, however, should provide the same result if the experimental correlation functions faithfully reproduce ensemble averages at the steady state.

This strategy cannot be adapted to interacting problems, outside of the mean field approximation. The obstacle comes from the parametrization of the interaction matrix, which is the discrete counterpart of introducing an interaction range in the corresponding field theory. Without *a priori* parametrization, the issue of sufficient statistics arises: One can think about repeating the same procedure in the multiparticle case for each particle pair and look for independent inference formulas for any matrix element $J\Lambda^{ij}$. Bypassing the technical difficulties related to solving the resulting system of N^2+2 second-degree equations for the unknowns b, $e^{-\eta\Delta t}$ and $\{J\Lambda^{ij}\}_{i,j=1\ldots N}$, we have a much greater number of parameters to infer than of points in each frame. This problem becomes totally untractable if one also allows Λ^{ij} to evolve in time, as in active animal groups [22,52].

Assumptions about the structure of the matrix Λ_{ij} dramatically diminish the number of parameters and help us deal with the worry of insufficient statistics but require an alternative strategy to estimate the interaction range, since this physically motivated parametrization does not allow us to find closed-form equations.

It is possible yet to approximately estimate the damping coefficient and the effective temperature of the system of interacting particles, assuming that they are all immersed in the same uniform thermal bath. Under this assumption, Eqs. (B46)–(B48) can be adapted to the interacting case and properly manipulated to find the following relations:

$$F_s - G'_s - G_s + C_s = e^{-(\eta/\chi)\Delta t}(2G'_s - C''_s - C_s) + \frac{G'_{\rm int} - C_{\rm int}}{C_{\rm int}}[G_s - C_s - b - e^{-(\eta/\chi)\Delta t}(C_s - G'_s)] - b; \tag{B55}$$

$$C'_s - 2G_s + C_s = e^{-(\eta/\chi)\Delta t}(G_s - F_s - C_s + G'_s) + b\{4 + e^{-(\eta/\chi)\Delta t} + \frac{n_c}{C_{\rm int}}[G_s - C_s - b - e^{-(\eta/\chi)\Delta t}(C_s - G'_s)]\} + \frac{G_{\rm int} - C_{\rm int}}{C_{\rm int}}[G_s - C_s - b - e^{-(\eta/\chi)\Delta t}(C_s - G'_s)]; \tag{B56}$$

where we use the third independent equation to eliminate J/χ and exploit the fact that $a = 4b$, with $b = \frac{1}{6}2T\eta\Delta t^3/\chi^2$. Let us define the empirical spatiotemporal correlation functions involved in these inference formulas.

(i) Equal-time correlations:

$$C_{ij} = \frac{1}{L-1}\sum_{n=1}^{L-1}\boldsymbol{\pi}_n^i \cdot \boldsymbol{\pi}_n^j; \tag{B57}$$

$$C'_{ij} = \frac{1}{L-1}\sum_{n=1}^{L-1}\boldsymbol{\pi}_{n+1}^i \cdot \boldsymbol{\pi}_{n+1}^j; \tag{B58}$$

$$C''_{ij} = \frac{1}{L-1}\sum_{n=1}^{L-1}\boldsymbol{\pi}_{n-1}^i \cdot \boldsymbol{\pi}_{n-1}^j. \tag{B59}$$

(ii) One-step correlations:

$$G_{ij} = \frac{1}{L-1}\sum_{n=1}^{L-1}\boldsymbol{\pi}_{n+1}^i \cdot \boldsymbol{\pi}_n^j; \tag{B60}$$

$$G'_{ij} = \frac{1}{L-1}\sum_{n=1}^{L-1}\boldsymbol{\pi}_n^i \cdot \boldsymbol{\pi}_{n-1}^j. \tag{B61}$$

(iii) Two-step correlations:

$$F_{ij} = \frac{1}{L-1}\sum_{n=1}^{L-1}\boldsymbol{\pi}_{n+1}^i \cdot \boldsymbol{\pi}_{n-1}^j. \tag{B62}$$

The observables appearing in Eqs. (B55) and (B56) are defined from Eqs. (B57)–(B62) as in the following. We can distinguish the contribution of self-correlations, encoded by

$$C_s = \frac{1}{N}\mathrm{Tr}\mathbf{C}; \qquad C'_s = \frac{1}{N}\mathrm{Tr}\mathbf{C}'; \qquad C''_s = \frac{1}{N}\mathrm{Tr}\mathbf{C}'';$$
$$G_s = \frac{1}{N}\mathrm{Tr}\mathbf{G}; \qquad G'_s = \frac{1}{N}\mathrm{Tr}\mathbf{G}'; \qquad F_s = \frac{1}{N}\mathrm{Tr}\mathbf{F},$$

and that of correlations between directly interacting birds, encoded by the quantities

$$C_{\rm int} = \frac{\mathrm{Tr}(\boldsymbol{\Lambda}\mathbf{C})}{N}; \quad G_{\rm int} = \frac{\mathrm{Tr}(\boldsymbol{\Lambda}^\top\mathbf{G})}{N}; \quad G'_{\rm int} = \frac{\mathrm{Tr}(\boldsymbol{\Lambda}\mathbf{G}')}{N},$$

where $\Lambda_{ij} = n_c\delta_{ij} - n_{ij}$. Notice that all of them are by definition self-averaging quantities, which obviously tend to be more and more stable as the size of the system increases.

As already stressed, in the absence of a proper likelihood, an unattainable task is that of dealing with functions denoted with an "int" subscript. However, the manipulation we carry out to derive Eqs. (B55) and (B56) confines them into subleading terms, as can be checked by looking at the combination

$$\frac{G_{\rm int} - C_{\rm int}}{C_{\rm int}}[(1 - e^{-(\eta/\chi)\Delta t})(G_s - C_s) - b] \simeq O(\Delta t^5), \tag{B63}$$

at the one obtained replacing $G_{\rm int}$ with $G'_{\rm int}$, and at

$$b \cdot \frac{n_c}{C_{\rm int}}[(1 - e^{-(\eta/\chi)\Delta t})(G_s - C_s) - b] \simeq O(\Delta t^6). \tag{B64}$$

Under the working hypothesis that Δt is sufficiently small, we can neglect these terms and find usable relations to extract the effective parameters of the thermal bath $(\eta/\chi,\ T/\chi)$ from the experimental self-correlations only. Precisely, η/χ is found as a solution of the equation

$$(C''_s + C_s - 2G_s)e^{-2(\eta/\chi)\Delta t} + 2(F_s - 5G'_s - G_s + 3C_s + 2C''_s)e^{-(\eta/\chi)\Delta t} + 4F_s - 4G'_s - 6G_s + 5C_s + C'_s = 0, \tag{B65}$$

whereas the effective temperature is extracted from b, being

$$b = G_s' + G_s - F_s - C_s + e^{-(\eta/\chi)\Delta t}(2G_s' - C_s - C_s''). \tag{B66}$$

Notice that this formula is exactly equivalent to Eq. (B53), since we define the effective damping coefficient of the harmonic oscillator as $\eta = \mu/m$, whereas the corresponding quantity, having the dimension of an inverse timescale, is η/χ for the ISM. These formulas are applied to find the results shown in Fig. 5.

APPENDIX C: EQUATIONS OF MOTION OF THE ISM IN THE SWA

We derive in this Appendix the equations of motion of the ISM in the so-called SWA. The name comes from the analogy with ideal Heisenberg ferromagnets which, at very low temperatures, can be studied using an approximate theory, whose basic idea is that the lowest-energy excitations in a ferromagnet are those produced by a single reversed spin over a large number of otherwise aligned spins in a crystal lattice [53]. In a similar way, since natural flocks of starlings are in a deeply ordered phase, we can perform an expansion around the perfectly ordered state of the flock, where all of the birds' velocities are aligned along the same direction.

Let us denote by $\mathbf{n}$ the collective direction of motion of the flock. Each vector $\mathbf{v}_i$ can be decomposed into its longitudinal and transverse components with respect to $\mathbf{n}$:

$$\mathbf{v}_i = v_i^L \mathbf{n} + \boldsymbol{\pi}_i. \tag{C1}$$

In the case of bird flocks, the spin-wave approximation reduces to approximating the longitudinal components as follows:

$$v_i^L = \sqrt{1 - |\boldsymbol{\pi}_i|^2} \simeq 1 - \frac{1}{2}|\boldsymbol{\pi}_i|^2, \tag{C2}$$

having $\mathbf{v}_i$ a unit length. The equations of motion of the ISM (with a fixed interaction network) can be written in the form of a set of second-order SDEs for the velocity variables:

$$\frac{d^2\mathbf{v}_i}{dt^2} = \left(-\eta\frac{d\mathbf{v}_i}{dt} - J\sum_{j=1}^{N} n_{ij}\mathbf{v}_j + \boldsymbol{\xi}_i\right)_\perp - \left|\frac{d\mathbf{v}_i}{dt}\right|^2 \mathbf{v}_i, \tag{C3}$$

where the $\perp$ symbol indicates the projection onto the orthogonal plane to the direction of motion of the ith bird, $\mathbf{v}_i$. This projection operator and the last term of Eq. (C3) are the required ingredients to ensure individual speed conservation: $|\mathbf{v}_i(t)| = v_0 = 1 \;\; \forall \;\; i, t$. Thanks to this property, Eq. (C3) further simplifies:

$$\frac{d^2\mathbf{v}_i}{dt^2} = -\eta\frac{d\mathbf{v}_i}{dt} - J\sum_{j=1}^{N} n_{ij}\mathbf{v}_{j\perp} + \boldsymbol{\xi}_{i\perp} - \left|\frac{d\mathbf{v}_i}{dt}\right|^2 \mathbf{v}_i. \tag{C4}$$

Using Eqs. (C1) and (C2) and exploiting the fact that, for any vector $\mathbf{a}$,

$$\mathbf{a}_\perp = -\mathbf{v}_i \times (\mathbf{v}_i \times \mathbf{a}), \tag{C5}$$

one can evaluate all the terms appearing in Eq. (C4), at the desired order of approximation.

Let us focus first on time derivatives: We notice that, in principle, they also produce terms containing $d\mathbf{n}/dt$ and $d^2\mathbf{n}/dt^2$. In the following, we assume that the direction of collective motion $\mathbf{n}$ is constant. This assumption is legitimate in the limit $N \to \infty$, when the wandering of the order parameter is suppressed, or at least when it is very slow compared to the relaxational dynamics of the degrees of freedom. If, on the contrary, one wants to take this effect into account, apparent forces emerge, because the chosen reference frame is noninertial.

Neglecting apparent forces enables one to segregate *on-plane* (i.e., perpendicular to $\mathbf{n}$) and *off-plane* (i.e., parallel to $\mathbf{n}$) contributions and completely disentangle the corresponding equations. One can then consider the equations in the $\boldsymbol{\pi}$ plane only:

$$\frac{d^2\boldsymbol{\pi}_i}{dt^2} + \eta\frac{d\boldsymbol{\pi}_i}{dt} + J\Lambda_{ij}\boldsymbol{\pi}_j = \hat{P}\boldsymbol{\xi}_{i\perp} + O(|\boldsymbol{\pi}|^3), \tag{C6}$$

where $\Lambda_{ij} = n_{ij} - n_c\delta_{ij}$ and $\hat{P}$ is the projection operator onto the plane perpendicular to the collective velocity $\mathbf{V} = (1/N)\sum_{i=1}^{N}\mathbf{v}_i \equiv \Phi\mathbf{n}$. The velocity fluctuations $\boldsymbol{\pi}_i$ play in this case the same role as spin excitations in Dyson's SWA, both becoming the new degrees of freedom and displaying a linear interaction.

At this stage, what remains to explicitly evaluate is only $\hat{P}\boldsymbol{\xi}_{i\perp}$. We know that $\boldsymbol{\xi}_{i\perp}$ lives in the plane perpendicular to $\mathbf{v}_i$, so that the perpendicular component to the plane spanned by $\mathbf{V}$ and $\mathbf{v}_i$ is left unchanged by this projection operator, while the other one is contracted with a factor $\cos\theta_i$, with θ_i the angle between $\mathbf{v}_i$ and $\mathbf{n}$. As a result,

$$\langle \hat{P}\boldsymbol{\xi}_i(t) \cdot \hat{P}\boldsymbol{\xi}_i(s)\rangle = 2(1 + \cos^2\theta_i)\frac{T\eta}{\chi^2}\delta(t - s). \tag{C7}$$

The second moment of each noise term is then rescaled, with respect to the original one, by a factor

$$\frac{1}{2}(1 + \cos^2\theta_i) = \frac{1}{2}[1 + (v_i^L)^2] = 1 - \frac{1}{2}|\boldsymbol{\pi}_i|^2 \simeq v_i^L. \tag{C8}$$

In order to let the fluctuation-dissipation theorem hold, this rescaling can be readsorbed by the temperature parameter T/χ, which is, in principle, different for each bird. At an averaged level, we can define a new spin-wave temperature that differs form the original temperature of the inertial spin model by a factor $(1/N)\sum_{i=1}^{N} v_i^L$, which is by definition equivalent to the polarization of the flock

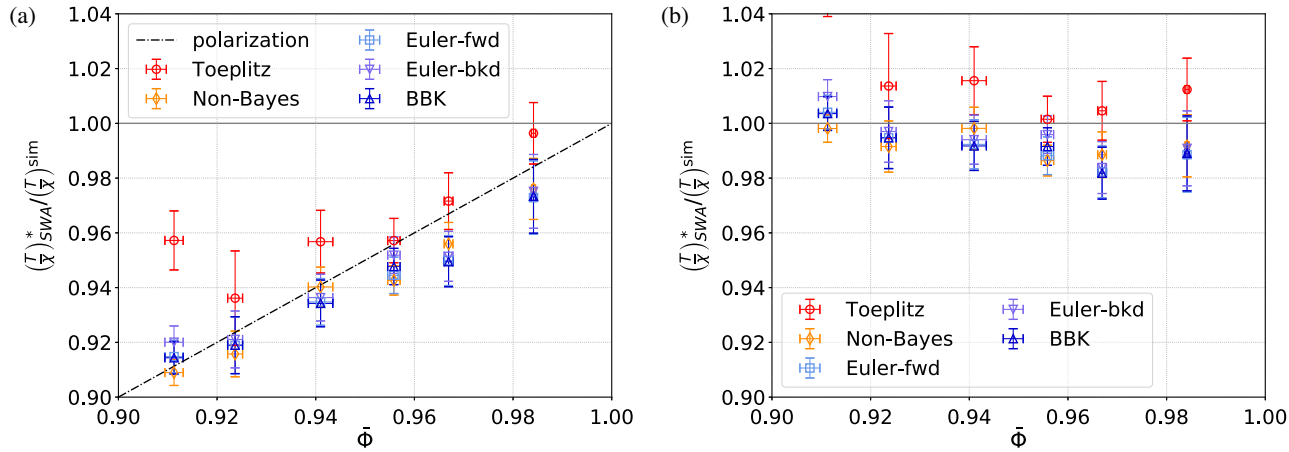

FIG. 9. First correction to the SWA. The comparison between the plots shows the effect of the SWA: In (a), the raw inferred values of T/χ, obtained using the inference formulas derived from Eq. (41), are reported. In (b), we include the first correction by rescaling the output with the time-averaged polarization, for each sample trajectory. $\bar{\Phi}$ is the average of the averaged polarizations among different simulated flocks, at any given temperature. Error bars for $\bar{\Phi}$ correspond to standard errors, whereas vertical bars represent, as in the other figures, 0.95 CI.

$\Phi = |(1/N)\sum_{i=1}^{N} \mathbf{v}_i|$. In the low-temperature case, where $|\boldsymbol{\pi}| \ll 1$, $\Phi = 1 + O(|\boldsymbol{\pi}|^2)$; the first correction to the temperature parameter is then of a lower order with respect to the terms which have been neglected in the deterministic part of Eq. (C6) and shall correctly be included through this simple effective rescaling.

As long as the experimental or statistical errors are wide enough and the system pretty ordered, this SWA-related correction is negligible. Thanks to the large statistics and high accuracy we manage to have with our simulations and inference machinery, we are able to detect it in Fig. 5(c), where points are systematically placed below the line of slope 1, especially for higher values of the temperature, which, in turn, correspond to lower polarization values. A comparison between the two panels in Fig. 9 confirms that this correction to the SWA is truly the origin of the observed trend and not an intrinsic defect of the inference procedure.

APPENDIX D: ISM SIMULATIONS

We implement a numerical integrator for the ISM in $d = 3$ that combines the leapfrog method with Boris's trick to ensure speed conservation [54]. We perform simulations on fixed Poisson random lattices (i.e., sites are randomly chosen points with uniform distribution), discarding the update of particle positions and consequent reshuffling effects. As a result, the adjacency matrix of the graph associated to the interacting particle system is time independent, and the constant speed v_0 of each bird does not play any role. Thus, the numerical integrator we use consists of the following set of update equations:

$$\mathbf{v}_i^{n+1} = \mathbf{v}_i^n + (\mathbf{v}_i^n + \mathbf{v}_i^n \times \mathbf{t}^n) \times \mathbf{u}^n, \qquad \mathbf{s}_i^{n+1/2} = \left(1 + \frac{\eta\Delta t}{2\chi}\right)^{-1} \left\{ \left(1 - \frac{\eta\Delta t}{2\chi}\right) \mathbf{s}_i^{n-1/2} + \mathbf{v}_i^n \times \left[\frac{J\Delta t}{\chi} \sum_j n_{ij} \mathbf{v}_j^n + \boldsymbol{\Xi}_i^n \right] \right\} \tag{D1}$$

with $\mathbf{t}^n = -(1/2\chi)\Delta t \mathbf{s}^{n+1/2}$ and $\mathbf{u}^n = 2\mathbf{t}^n/(1 + |\mathbf{t}^n|^2)$. $\boldsymbol{\Xi}_i^n$ is a three-dimensional isotropic Gaussian variable of zero mean and of variance

$$\langle \boldsymbol{\Xi}_i^n \cdot \boldsymbol{\Xi}_j^m \rangle = \delta_{ij}\delta_{mn} 2 \cdot 3 \cdot T\eta\Delta t. \tag{D2}$$

The adjacency matrix explicitly reads

$$n_{ij} = \begin{cases} 1 & \text{if } r_{ij} \le n_c, \\ 0 & \text{if } r_{ij} > n_c \end{cases} \tag{D3}$$

with r_{ij} the rank of bird j as a neighbor of bird i (excluding the bird itself, to which we conventionally associate rank $r_{ii} = 0$). In all of our simulations, we work with periodic boundary conditions.

We try to ensure that the system is sampled in a stationary regime by starting from microscopic configurations corresponding to polarization values close to the equilibrium ones. The polarization is the macroscopic order parameter of the system, and it is defined, in perfect analogy to the magnetization in a three-dimensional Heisenberg model, as $\Phi = (1/Nv_0)|\sum_{i=1}^{N} \mathbf{v}_i|$.

Flocks of $N = 1000$ birds are simulated to obtain the results shown in this paper, with a topological range of interaction $n_c = 6$ [except for the data in Fig. 5(b)], alignment strength $J/\chi = 5$, and effective temperature T/χ in the range [0.2, 1.2]. When not explicitly indicated, we take $T/\chi = 0.4$, approximately corresponding to a polarization of 0.97 (for $n_c = 6$). We choose an integration time step of $\tau_{\rm sim} = 0.0005/(Jn_c)$ for all the simulations. Different damping regimes are explored, and the performance of the inference method is tested in each of them and for various choices of the time lag Δt. In order to disentangle the effects of the discrete nature of the simulation from proper malfunctioning of the inference schemes, the minimum inference time step Δt displayed in Figs. 5(b) and 8 is $5\tau_{\rm sim}$.

[1] D. B. Brückner, A. Fink, C. Schreiber, P. J. F. Röttgermann, J. O. Rädler, and C. P. Broedersz, *Stochastic Nonlinear Dynamics of Confined Cell Migration in Two-State Systems*, Nat. Phys. **15**, 595 (2019).

[2] F. Gerbal, P. Chaikin, Y. Rabin, and J. Prost, *An Elastic Analysis of Listeria Monocytogenes Propulsion*, Biophys. J. **79**, 2259 (2000).

[3] G. J. Stephens, B. Johnson-Kerner, W. Bialek, and W. S. Ryu, *Dimensionality and Dynamics in the Behavior of C. Elegans*, PLoS Comput. Biol. **4**, e1000028 (2008).

[4] J. Gautrais, C. Jost, M. Soria, A. Campo, S. Motsch, R. Fournier, S. Blanco, and G. Theraulaz, *Analyzing Fish Movement as a Persistent Turning Walker*, J. Math. Biol. **58**, 429 (2009).

[5] J. Gautrais, F. Ginelli, R. Fournier, S. Blanco, M. Soria, H. Chaté, and G. Theraulaz, *Deciphering Interactions in Moving Animal Groups*, PLoS Comput. Biol. **8**, e1002678 (2012).

[6] A. Attanasi, A. Cavagna, L. Del Castello, I. Giardina, T. S. Grigera, A. Jelić, S. Melillo, L. Parisi, O. Pohl, E. Shen, and M. Viale, *Information Transfer and Behavioural Inertia in Starling Flocks*, Nat. Phys. **10**, 691 (2014).

[7] A. Cavagna, D. Conti, C. Creato, L. Del Castello, I. Giardina, T. S. Grigera, S. Melillo, L. Parisi, and M. Viale, *Dynamic Scaling in Natural Swarms*, Nat. Phys. **13**, 914 (2017).

[8] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University, New York, 2001).

[9] M. S. Miguel and J. M. Sancho, *A Colored-Noise Approach to Brownian Motion in Position Space. Corrections to the Smoluchowski Equation*, J. Stat. Phys. **22**, 605 (1980).

[10] P. Hanggi, H. Thomas, H. Grabert, and P. Talkner, *Note on Time Evolution of Non-Markov Processes*, J. Stat. Phys. **18**, 155 (1978).

[11] D. B. Brückner, P. Ronceray, and C. P. Broedersz, *Inferring the Non-linear Dynamics of Stochastic Inertial Systems*, arXiv:2002.06680.

[12] B. Lehle and J. Peinke, *Analyzing a Stochastic Time Series Obeying a Second-Order Differential Equation*, Phys. Rev. E **91**, 062113 (2015).

[13] B. Lehle and J. Peinke, *Analyzing a Stochastic Process Driven by Ornstein-Uhlenbeck Noise*, Phys. Rev. E **97**, 012113 (2018).

[14] J. N. Pedersen, L. Li, C. Grădinaru, R. H. Austin, E. C. Cox, and H. Flyvbjerg, *How to Connect Time-Lapse Recorded Trajectories of Motile Microorganisms with Dynamical Models in Continuous Time*, Phys. Rev. E **94**, 062401 (2016).

[15] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill, *Principles of Maximum Entropy and Maximum Caliber in Statistical Physics*, Rev. Mod. Phys. **85**, 1115 (2013).

[16] H. C. Nguyen, R. Zecchina, and J. Berg, *Inverse Statistical Problems: From the Inverse Ising Problem to Data Science*, Adv. Phys. **66**, 197 (2017).

[17] D. J. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, England, 2003).

[18] M. Sørensen, *Estimating Functions for Discretely Observed Diffusions: A Review, in Selected Proceedings of the Symposium on Estimating Functions*, IMS Lecture Notes Monograph Series Vol. 32, edited by V. P. G. I. V. Basawa and R. L. Taylor (Institute of Mathematical Statistics, Hayward, CA, 1997), pp. 305–325.

[19] C. Monthus, *Non-equilibrium Steady States: Maximization of the Shannon Entropy Associated with the Distribution of Dynamical Trajectories in the Presence of Constraints*, J. Stat. Mech. (2011) P03008.

[20] M. Otten and G. Stock, *Maximum Caliber Inference of Nonequilibrium Processes*, J. Chem. Phys. **133**, 034119 (2010).

[21] A. Cavagna, I. Giardina, F. Ginelli, T. Mora, D. Piovani, R. Tavarone, and A. M. Walczak, *Dynamical Maximum Entropy Approach to Flocking*, Phys. Rev. E **89**, 042707 (2014).

[22] T. Mora, A. M. Walczak, L. Del Castello, F. Ginelli, S. Melillo, L. Parisi, M. Viale, A. Cavagna, and I. Giardina, *Local Equilibrium in Bird Flocks*, Nat. Phys. **12**, 1153 (2016).

[23] A. Frishman and P. Ronceray, *Learning Force Fields from Stochastic Trajectories*, Phys. Rev. X **10**, 021009 (2020).

[24] M. El Beheiry, S. Türkcan, M. U. Richly, A. Triller, A. Alexandrou, M. Dahan, and J.-B. Masson, *A Primer on the Bayesian Approach to High-Density Single-Molecule Trajectories Analysis*, Biophys. J. **110**, 1209 (2016).

[25] S. Türkcan, A. Alexandrou, and J.-B. Masson, *A Bayesian Inference Scheme to Extract Diffusivity and Potential Fields from Confined Single-Molecule Trajectories*, Biophys. J. **102**, 2288 (2012).

[26] A. Gloter, *Parameter Estimation for a Discretely Observed Integrated Diffusion Process*, Scand. J. Stat. Theory Appl. **33**, 83 (2006).

[27] A. Gloter, *Parameter Estimation for a Discrete Sampling of an Integrated Ornstein-Uhlenbeck Process*, Statistics **35**, 225 (2001).

[28] We seek a solution of the stochastic process either as an explicit sample-path solution in position space that does not involve integro-differential operators or as the time-dependent solution of the associated Fokker-Planck equation [29].

[29] C. Gardiner, *Stochastic Methods* (Springer-Verlag, Berlin, 2009), Vol. 13.

[30] A. N. Drozdov and M. Morillo, *Path Integral Solution of the Kramers Problem*, Phys. Rev. Lett. **77**, 5324 (1996).

[31] P. E. K. E. Platen and P. E. Kloeden, *Numerical Solution of Stochastic Differential Equations*, 1st ed., Stochastic Modelling and Applied Probability Vol. 23 (Springer-Verlag, Berlin, 1992).

[32] D. T. Gillespie, *Exact Numerical Simulation of the Ornstein-Uhlenbeck Process and Its Integral*, Phys. Rev. E **54**, 2084 (1996).

[33] A. Cavagna, D. Conti, I. Giardina, T. S. Grigera, S. Melillo, and M. Viale, *Spatio-Temporal Correlations in Models of Collective Motion Ruled by Different Dynamical Laws*, Phys. Biol. **13**, 065001 (2016).

[34] A. Cavagna, I. Giardina, and T. S. Grigera, *The Physics of Flocking: Correlation as a Compass from Experiments to Theory*, Phys. Rep. **728**, 1 (2018).

[35] The class of models considered by Gloter in Ref. [26] is not exactly the same as the one we consider in Eq. (1) [$f(x) = 0$ is assumed and the presence of nonlinear nonconservative forces and of a multiplicative noise of the form of $\sigma(v)\xi$ is allowed], but we think that the result in Ref. [26] may be extended also to the $f(x) \neq 0$ case.

[36] This result is not surprising if one carefully looks at the expression of the inverse of the tridiagonal Toeplitz matrix Eq. (29), which closely resembles Fourier series expansions. Increasing the number of points corresponds to including an increasing number of harmonics;finite-size corrections to parameters estimators can be seen as a counterpart of the Gibbs phenomenon.

[37] J. Jia, T. Sogabe, and M. El-Mikkawy, *Inversion of k-Tridiagonal Matrices with Toeplitz Structure*, Comput. Math. Appl. **65**, 116 (2013).

[38] G. Meurant, *A Review on the Inverse of Symmetric Tridiagonal and Block Tridiagonal Matrices*, SIAM J. Matrix Anal. Appl. **13**, 707 (1992).

[39] E. Vanden-Eijnden and G. Ciccotti, *Second-Order Integrators for Langevin Equations with Holonomic Constraints*, Chem. Phys. Lett. **429**, 310 (2006).

[40] R. D. Skeel and J. A. Izaguirre, *An Impulse Integrator for Langevin Dynamics*, Mol. Phys. **100**, 3885 (2002).

[41] W. van Gunsteren and H. Berendsen, *Algorithms for Brownian Dynamics*, Mol. Phys. **45**, 637 (1982).

[42] There exists a similarity transformation that transforms the matrix in Eq. (35) into a strictly diagonally dominant matrix with positive entries. Since the spectrum is unchanged and it is real, the non-negativity of all the eigenvalues is ensured.

[43] A. Dequidt and J. G. Solano Canchaya, *Bayesian Parametrization of Coarse-Grain Dissipative Dynamics Models*, J. Chem. Phys. **143**, 084122 (2015).

[44] V. Harmandaris, E. Kalligiannaki, M. Katsoulakis, and P. Plecháč, *Path-Space Variational Inference for Non-equilibrium Coarse-Grained Systems*, J. Comput. Phys. **314**, 355 (2016).

[45] A. Cavagna, L. Del Castello, I. Giardina, T. Grigera, A. Jelic, S. Melillo, T. Mora, L. Parisi, E. Silvestri, M. Viale, and A. M. Walczak, *Flocking and Turning: A New Model for Self-Organized Collective Motion*, J. Stat. Phys. **158**, 601 (2015).

[46] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic, *Interaction Ruling Animal Collective Behaviour Depends on Topological Rather Than Metric Distance: Evidence from a Field Study*, Proc. Natl. Acad. Sci. U.S.A. **105**, 1232 (2008).

[47] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, *Statistical Mechanics for Natural Flocks of Birds*, Proc. Natl. Acad. Sci. U.S.A. **109**, 4786 (2012).

[48] A. Cavagna, L. Di Carlo, I. Giardina, L. Grandinetti, T. S. Grigera, and G. Pisegna, *Renormalization Group Crossover in the Critical Dynamics of Field Theories with Mode Coupling Terms*, Phys. Rev. E **100**, 062130 (2019).

[49] A. Cavagna, L. Di Carlo, I. Giardina, L. Grandinetti, T. S. Grigera, and G. Pisegna, *Dynamical Renormalization Group Approach to the Collective Behavior of Swarms*, Phys. Rev. Lett. **123**, 268001 (2019).

[50] R. Mannella and V. Palleschi, *Fast and Precise Algorithm for Computer Simulation of Stochastic Differential Equations*, Phys. Rev. A **40**, 3381 (1989).

[51] A. Brünger, C. L. Brooks, and M. Karplus, *Stochastic Boundary Conditions for Molecular Dynamics Simulations of st2 Water*, Chem. Phys. Lett. **105**, 495 (1984).

[52] A. Cavagna, S. M. D. Queirós, I. Giardina, F. Stefanini, and M. Viale, *Diffusion of Individual Birds in Starling Flocks*, Proc. R. Soc. B **280**, 20122484 (2013).

[53] F. J. Dyson, *General Theory of Spin-Wave Interactions*, Phys. Rev. **102**, 1217 (1956).

[54] J. P. Boris, *Relativistic Plasma Simulation-Optimization of a Hybrid Code*, in *Proceedings of the 4th Conference on Numerical Simulation of Plasmas 3* (1970), https://ci.nii.ac.jp/naid/10009996893/en/.