Statistics Surveys
Vol. 9 (2015) 209–252
ISSN: 1935-7516
DOI: 10.1214/15-SS111

Statistical inference for dynamical systems: A review

Kevin McGoff

e-mail: kmcgoff1@uncc.edu

Sayan Mukherjee

e-mail: sayan@stat.duke.edu

and

Natesh Pillai

e-mail: pillai@fas.harvard.edu

Abstract: The topic of statistical inference for dynamical systems has been studied widely across several fields. In this survey we focus on methods related to parameter estimation for nonlinear dynamical systems. Our objective is to place results across distinct disciplines in a common setting and highlight opportunities for further research.

MSC 2010 subject classifications: Primary 62M09; secondary 60G35, 37A50.
Keywords and phrases: Parameter estimation, consistency, dynamical systems, ergodic theory.

Received February 2014.

1. Introduction

The problem of parameter estimation in dynamical systems appears in many areas of science and engineering. Often the form of the model can be derived from some knowledge about the process under investigation, but parameters of the model must be inferred from empirical observations in the form of time series data. As this problem has appeared in many different contexts, partial solutions have been proposed in a wide variety of disciplines, including nonlinear dynamics in physics, control theory in engineering, state space modeling in statistics and econometrics, and ergodic theory and dynamical systems in mathematics. One purpose of this study is to present these various approaches in a common language, with the hope of unifying some ideas and pointing towards interesting avenues for further study. We focus here on theoretical results and methods for inference, as a detailed presentation of applied work and example data sets lies outside the scope of this survey.

We will concern ourselves with stochastic processes of the form $(X_t, Y_t)_t$, where t runs over either $\mathbb{R}_+$ (continuous-time) or $\mathbb{Z}_+$ (discrete-time). In order

to treat several cases, including examples such as ordinary differential equations, with a single formalism, we allow either one or both of $(X_t)_t$ and $(Y_t)_t$ to be "deterministic," in a sense made precise in Section 2. In the discrete-time setting, we may write $(X_n, Y_n)_n$. We think of X_t as the true state of the system at time t and Y_t as an observation of the system at time t.

The case when no noise is present has been most often considered by mathematicians in the field of dynamical systems and ergodic theory. In this case, all uncertainty in the system comes from the uncertainty in the initial state of the system, and the ability to estimate any parameters in the system may depend strongly on properties of the observation function $f(X_t) = Y_t$. State space models, considered most often by statisticians, lie at the other end of the noise spectrum, where both X_t and Y_t depend on some noise. Hidden Markov models, which have received considerable attention, provide a broad class of examples of these systems. In this setting, the statistical question of consistency for methods of parameter estimation has been studied, and some general results are available. The other two possible assumptions on the presence of noise (only dynamical noise or only observational noise) have received relatively little attention, especially from the statistical point of view. Let us mention that very often both dynamical noise and observational noise arise in real data examples, especially in biology and ecology [15, 35, 91, 96, 107, 164, 173]; nonetheless, as the other settings arise in some physical models, we discuss each of the logical possibilities for the noise structure.

As an example of the type of questions of interest, consider the question of parameter inference for models of gene regulatory networks [5, 151, 162, 171]. The underlying model often favored by biologists consists of a system of ordinary differential equations, with each variable in the state vector representing the expression level of a particular gene in the network [5, 171]. For some networks of interest, a significant amount of work has produced biological understanding regarding the qualitative interactions between the genes in the network, but the corresponding ODE models still contain several parameters necessary for quantifying these interactions [32, 151]. Experimentalists are able to conduct experiments in which the expression levels of the genes in the network are measured at regularly spaced instances of time. The resulting data may be interpreted as time series data generated by a system of ODEs with noisy observations. The parameter inference problem in this setting consists of inferring the parameters of the ODE model from the observed data, and to the best of our knowledge there are still significant statistical challenges in this area [140].

Another example of interest involves identifying the behavior of a dynamical system on a computer network. In a variety of applications one considers nodes in a communication network and measures the states of these nodes (or properties of the nodes) over time. In many settings, one would like to detect drastic changes in the nature of the dynamic behavior of the system. This problem is of vital importance to a variety of reliability and security applications on networks [64, 88, 98, 132] and it can be formalized as the inference of large changes in the parameters of the network – a change point model for a dynamic network.

The objective of this article is to survey methodology across a variety of fields

for parameter inference in dynamical systems. We first state the various goals of inference in dynamical systems. Our focus will be parameter inference, and we provide a natural classification of parameter inference into four possible settings defined by the structure of noise in the system. We then state what is known in terms of rigorous results for parameter inference in these four settings. Of these settings the case of deterministic dynamics with observational noise appears to be the least developed in terms of sound statistical theory. We also mention several important open problems for parameter inference.

There is an extremely large body of work stretching across many disciplines that relates to the topic of statistical properties of dynamical systems. Although we attempt to provide references when possible, we make no attempt to be exhaustive, and we recognize that in fact many references have been omitted. On the other hand, we hope that the references cited in this article may serve as an appropriate starting point for further reading.

1.1. Goals of statistical inference

There are a variety of topics that can be considered part of "statistical inference in dynamical systems." In the interest of providing context for this survey, let us mention the following topics:

1. parameter estimation, model identification or reconstruction;
2. state estimation, filtering, smoothing, or denoising;
3. feature estimation, where features often include invariant measures, dimensions, entropy, or Lyapunov exponents;
4. prediction or forecasting;
5. noise quantification, estimation, or detection.

In this paper we focus almost exclusively on the problems of parameter inference, system identification and reconstruction. Informally, we pose the parameter estimation problem as follows. Suppose the family of processes $(X_t, Y_t)_t$ under consideration can be parametrized by a set of parameters $\mathcal{A}$, with a serving as the "true" parameter controlling the system. Construct statistical procedures for estimating the parameter a, given observations $Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}$, and provide adequate theoretical support for the validity of the estimation procedure. This problem is often considered to have several components. First, one would like to perform an identifiability analysis to understand in what sense, if any, the parametrization is identifiable. Second, one would like to construct asymptotically consistent estimation procedures. Third, one would like to study finite sample properties and perhaps provide interval estimates to quantify uncertainty. Lastly, one would like to understand how these results depend on the modelling assumptions, including issues such as model misspecification.

Of course, the boundaries between the problems listed above are often quite blurred. For example, if one can accurately estimate the hidden states $(X_{t_k})_{k=0}^{n-1}$ from the data $(Y_{t_k})_{k=0}^{n-1}$, then the problem of system identification often becomes significantly easier. For this reason, parameter inference methods may simultaneously attempt some version of state estimation or denoising.

1.2. Organization of the paper

We organize this survey according to the structure of the noise in the system. This organization is motivated by the observation that methods and results for parameter inference in dynamical systems tend to be specific to the type of noise assumed in the model.

The remainder of the paper is organized as follows. In Section 2 we give some essential definitions in the field of dynamical systems and make some general statements regarding parameter inference that hold under any noise assumptions. In Section 3 we describe some results relevant to inference for dynamical systems in the absence of noise. Section 4 contains a variety of proposed methods dealing with the case of dynamical systems contaminated by observational noise only. Section 5 deals with the case of only dynamical noise, and Section 6 addresses the setting of general state space models–that is, systems with both dynamical and observational noise. Lastly, we highlight some interesting open questions in Section 7.

Ornstein and Weiss [127] have shown that in a certain sense it is impossible, in general, to tell the difference between observational and dynamical noise. In this sense, one might suggest that from the point of view of abstract ergodic theory, we should not make distinctions on the basis of the type of noise present. However, we are often interested in finer properties than those captured by the equivalence relations considered in [127], and therefore the distinction between observational and dynamical noise is still useful for our purposes.

1.3. Related surveys and books

There have been many other reviews related to the topics in this survey. An incomplete list of such reviews is the following: [11, 18, 26, 49, 69, 73, 92, 156, 167]. Furthermore, let us mention the following books or monographs related to the topics in this survey: [1, 14, 25, 48, 86, 89, 127, 161]. The relevance of this survey is that we bring together approaches from many distinct fields and discuss them in a common statistical setting. In particular we discuss parameter estimation and inference for the full range of noise settings. This perspective is rare, since the different noise settings often correspond to different research areas, such as deterministic dynamics or state space methods based on hidden Markov models. We bring these various approaches together and place them in a common context.

2. Basic definitions and preliminaries

The most general setting that we will consider may be described as follows. Let $\mathcal{A}$, $\mathcal{X}$, and $\mathcal{Y}$ be Polish spaces [1] (complete metric spaces with a countable dense set), where each one is equipped with its Borel σ-algebra. The space $\mathcal{A}$ denotes

[1] This assumption is standard in dynamical systems and probability theory.

the parameter space, the underlying dynamical system evolves in the space $\mathcal{X}$, called the phase space or the state space, and the observations take values in $\mathcal{Y}$. For each a in $\mathcal{A}$, we consider a stochastic process $(X_t, Y_t)_t$ taking values in $\mathcal{X} \times \mathcal{Y}$ with law $\mathbb{P}_a$ such that for $t_0 < \cdots < t_n < t_{n+1}$, we have

$$\mathbb{P}_a(X_{t_{n+1}} \mid X_{t_0}, \ldots, X_{t_n}, Y_{t_0}, \ldots, Y_{t_n}) = \mathbb{P}_a(X_{t_{n+1}} \mid X_n) \tag{2.1}$$

$$\mathbb{P}_a(Y_{t_n} \mid X_{t_0}, \ldots, X_{t_n}, Y_{t_0}, \ldots, Y_{t_{n-1}}) = \mathbb{P}_a(Y_{t_n} \mid X_{t_n}). \tag{2.2}$$

We refer to the process $(X_t)_t$ as the underlying system (or trajectory) and the process $(Y_t)_t$ as the observation process. Note that the processes $(X_t)_t$ and $(X_t, Y_t)_t$ are both Markov, while the process $(Y_t)_t$ is not Markov in general.

We now distinguish between the four possible noise regimes. Let Var_a denote the variance operator with respect to the measure $\mathbb{P}_a$. We say the process $(X_t, Y_t)_t$ has dynamical noise if for some $s < t$ we have that

$$\mathrm{Var}_a(X_t \mid X_s) > 0.$$

We say the process $(X_t, Y_t)_t$ has observational noise if for some $t > 0$ we have that

$$\mathrm{Var}_a(Y_t \mid X_t) > 0.$$

Thus, the four possible noise settings are: no noise, observational noise only, dynamical noise only, or both dynamical and observational noise. Note that the Markov structure of the system is particularly relevant in the presence of dynamical noise (see Sections 5 and 6). Without dynamical noise, the fact that the process $(X_t)_t$ can be written as a Markov chain is less relevant, since the lack of variance makes it an especially degenerate chain. Thus, in the absence of dynamical noise, the theory of (deterministic) dynamical systems plays a more significant role.

We will have need to refer to stationary stochastic processes, which we define here.

Definition 2.1. A stochastic process $(X_t)_t$ is stationary if for any $k \in \mathbb{N}$, $t > 0$ and $t_1, \ldots, t_k$, the joint distribution of $(X_{t_1+t}, \ldots, X_{t_k+t})$ is equal to the joint distribution of $(X_{t_1}, \ldots, X_{t_k})$.

Let us mention that there is a natural correspondence between stationary stochastic processes and dynamical systems (see Remark 2.9 in Section 2.1 for a discussion of this correspondence in the discrete-time setting).

2.1. Remarks on discrete-time systems

Consider a process $(X_n, Y_n)_n$ satisfying Equations (2.1)–(2.2), with n in $\mathbb{Z}_+$.

Definition 2.2. An $\mathcal{X}$-valued stationary stochastic process $(X_n)_n$ is said to be ergodic if for every $\ell \geq 1$ and every pair of Borel sets $A, B \in \mathcal{X}^\ell$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{P}\Big((X_1, \ldots, X_\ell) \in A,\, (X_{k+1}, \ldots, X_{k+\ell}) \in B\Big)$$

$$= \mathbb{P}\Big((X_1, \ldots, X_\ell) \in A\Big)\mathbb{P}\Big((X_1, \ldots, X_\ell) \in B\Big).$$

Example 2.3. Many familiar processes are ergodic, including any i.i.d. process. Indeed, if $(X_n)_n$ is i.i.d., then for any k and ℓ, we have

$$\begin{aligned}
&\mathbb{P}\Big((X_1, \ldots, X_\ell) \in A, (X_{k+1}, \ldots, X_{k+\ell}) \in B\Big) \\
&\quad = \mathbb{P}\Big((X_1, \ldots, X_\ell) \in A\Big)\mathbb{P}\Big((X_{k+1}, \ldots, X_{k+\ell}) \in B\Big) \\
&\quad = \mathbb{P}\Big((X_1, \ldots, X_\ell) \in A\Big)\mathbb{P}\Big((X_1, \ldots, X_\ell) \in B\Big),
\end{aligned}$$

which shows that $(X_n)_n$ is ergodic.

Definition 2.4. A measurable dynamical system is a triple $(\mathcal{X}, \mathcal{F}, T)$, where $(\mathcal{X}, \mathcal{F})$ is a measurable space and $T : \mathcal{X} \to \mathcal{X}$ is measurable. A topological dynamical system is a pair $(\mathcal{X}, T)$, where $\mathcal{X}$ is a topological space and $T : \mathcal{X} \to \mathcal{X}$ is a continuous map. In the study of topological dynamics, one often assumes that $\mathcal{X}$ is compact and metrizable.

Example 2.5. Although there are many more elaborate examples of dynamical systems, let us consider here a basic family of examples, called circle rotations. Let $\mathcal{X}$ be the unit interval $[0, 1]$ with the endpoints identified, making it a topological circle, and let $\mathcal{F}$ be the σ-algebra of Borel measurable sets on $\mathcal{X}$. For any α in $\mathbb{R}$, let $R_\alpha : \mathcal{X} \to \mathcal{X}$ be defined by $R_\alpha(x) = x + \alpha \mod 1$, meaning that $R_\alpha(x)$ is the fractional part of $x + \alpha$. Then $(\mathcal{X}, \mathcal{F}, R_\alpha)$ forms a measurable dynamical system and $(\mathcal{X}, R_\alpha)$ forms a topological dynamical system. Under this system, points are rotated around the circle by an angle of $2\pi\alpha$ at each time step.

Definition 2.6. A measure-preserving system is a quadruple $(\mathcal{X}, \mathcal{F}, T, \mu)$, where $(\mathcal{X}, \mathcal{F}, \mu)$ is a measure space, $T : \mathcal{X} \to \mathcal{X}$ is measurable, and $\mu\big(T^{-1}(A)\big) = \mu(A)$ for each A in $\mathcal{F}$. In this case, we say that T preserves the measure μ and μ is an invariant measure for T. For the purpose of this article, we will always assume that any invariant measure μ is a probability measure. Also, if $\mathcal{X}$ is Polish and $\mathcal{F}$ is the Borel σ-algebra, then we may refer to $(\mathcal{X}, T, \mu)$ as a measure-preserving system.

Definition 2.7. A measure-preserving system $(\mathcal{X}, \mathcal{F}, T, \mu)$ is ergodic if whenever $T^{-1}(A) = A$ up to sets of μ-measure zero for A in $\mathcal{F}$, it happens that $\mu(A) \in \{0, 1\}$. We may say that T is ergodic for μ, or we may say that μ is ergodic for T.

Example 2.8. Consider again the circle rotations from Example 2.5. Let μ be Lebesgue measure on $\mathcal{X}$. Then $(\mathcal{X}, \mathcal{F}, R_\alpha, \mu)$ is a measure-preserving system, which can be seen by observing that R_α preserves the length and thus the measure of each interval in $\mathcal{X}$. Furthermore, one may check that $(\mathcal{X}, \mathcal{F}, R_\alpha, \mu)$ is ergodic if and only if α is irrational.

Remark 2.9. With the definitions given above, there is a correspondence between stationary stochastic processes and measure-preserving systems. Let us describe this correspondence as follows. Suppose $(X_n)_n$ is an $\mathcal{X}$-valued stationary stochastic sequence, where $\mathcal{X}$ is Polish. Let $\mathcal{Y} = \prod_n \mathcal{X}$, equipped with the product σ-algebra induced by the Borel σ-algebra on $\mathcal{X}$. Define $T : \mathcal{Y} \to \mathcal{Y}$ by the left shift: if $y = (x_n)_n$, then $(T(y))_n = x_{n+1}$. Kolmogorov's consistency theorem gives that there is a unique probability measure μ on $\mathcal{Y}$ with the same finite dimensional distributions as $(X_n)_n$. In this case, the stationarity of $(X_n)_n$ corresponds exactly to the invariance of μ with respect to T. Moreover, if $(X_n)_n$ is ergodic, then μ is ergodic for T.

In the other direction, given any measure-preserving system $(\mathcal{X}, T, \mu)$, we may define a stationary stochastic process as follows. For any Polish space $\mathcal{Y}$ and measurable map $f : \mathcal{X} \to \mathcal{Y}$, let $X_n(\omega) = f(T^n(\omega))$. If $(\mathcal{X}, T, \mu)$ is ergodic, then so is $(X_n)_n$.

Remark 2.10. In the discrete-time setting, any type of asymptotic analysis necessarily relies on the limit as the number of observations (n) tends to infinity. This limit, which involves observing the system over arbitrarily long time intervals, is often referred to as "out fill" asymptotics in the statistics literature.

Remark 2.11. Outside of the setting of finite state hidden Markov models, there seems to have been relatively little attention paid to the question of identifiability in the discrete-time setting. One could make the following definition. Let $\pi : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ be the projection onto the second coordinate. Then $\mathbb{P}_a \circ \pi^{-1}$ gives the law of the process $(Y_n)_n$. We say that the parametrization is identifiable if $\mathbb{P}_a \circ \pi^{-1} = \mathbb{P}_{a'} \circ \pi^{-1}$ implies $a = a'$. In the case when the parametrization is not identifiable, we could consider the problem of estimation up to the equivalence relation given by $a \sim a'$ whenever $\mathbb{P}_a \circ \pi^{-1} = \mathbb{P}_{a'} \circ \pi^{-1}$. We do not know of any general treatment of identifiability (in this or any other sense) for discrete-time systems.

2.2. Remarks on continuous-time systems

In the continuous-time setting, we restrict attention to the study of ordinary differential equations (ODEs) and stochastic differential equations (SDEs), allowing for the possibility of observational noise. Although many of the same ideas and difficulties are relevant to statistical inference of partial differential equations (see, for example, [175]), such systems lie outside the scope of the survey. Here we make a few remarks concerning these continuous-time systems.

Remark 2.12. Given a continuous-time system $(X_t, Y_t)_t$ and an interval length Δt, one can associate a discrete-time system as follows. Let $X'_n = X_{n\Delta t}$ and $Y'_n = Y_{n\Delta t}$, and then $(X'_n, Y'_n)_n$ is a discrete-time system as above. Unfortunately, if the system does not have a closed form solution (as is typically the case for nonlinear differential equations) and one is interested in parameter estimation, then the associated discrete-time process will not be explicitly parametrized. Thus, it seems necessary to develop methods of parameter estimation that are particular to the case of ODEs and SDEs.

Remark 2.13. There are two possible asymptotic regimes that may be considered in the study of inference for differential equations: "in fill" asymptotics, and "out fill" (also known as "expanding") asymptotics. In the case of "in fill" asymptotics, one prescribes a fixed interval of time on which the process will be observed, say $[a, b]$, and then one allows the number of observations in that time interval to grow to infinity (often with conditions on the distribution of sample times within the interval). In the case of "out fill" asymptotics, one obtains observations sequentially according to some sampling scheme on arbitrarily large (growing) time intervals.

Example 2.14. Here we discuss the system of ODEs that gives rise to the Lorenz attractor, which is a basic example of chaotic behavior. In the process of studying certain physical equations involved in modeling the weather, Lorenz made several simplifications and arrived at the following system of equations:

$$\begin{aligned}\frac{dx}{dt} &= \sigma(y - x)\\ \frac{dy}{dt} &= x(\rho - z) - y\\ \frac{dz}{dt} &= xy - \beta z,\end{aligned}$$

where the state of the system at time t is given by $(x(t), y(t), z(t))$ and σ, ρ, and β are physical parameters. In particular, Lorenz chose the parameters $\sigma = 10, \rho = 28$, and $\beta = 8/3$. With these parameter values, the system exhibits sensitivity to initial conditions, which means that if the system is started at two distinct positions, which might be arbitrarily close to each other, then the corresponding trajectories will eventually diverge from each other by a substantial amount. Such behavior is considered an indication of "chaos" in the system and is a fundamental property of many dynamical systems. See Figure 1 for a sample trajectory.

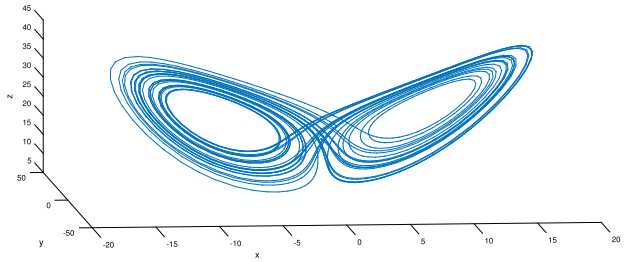

FIG 1. *A sample trajectory from the Lorenz system with classical parameters*

Example 2.15. Here we discuss mass-action stochastic kinetic models, a system of ODEs that appears in applications across statistics [19], applied probability [7], physical chemistry [158], and systems biology [170]. Given a set of chemical

species we consider the following set of reactions over the k species $X_1, \ldots, X_n$ and r reactions $\mathcal{R}_1, \ldots, \mathcal{R}_r$:

$$\begin{aligned}
\mathcal{R}_1 &: p_{11}X_1 + p_{12}X_2 + \cdots + p_{1N}X_n \to q_{11}X_1 + q_{12}X_2 + \cdots + q_{1k}X_k \\
\mathcal{R}_2 &: p_{21}X_1 + p_{22}X_2 + \cdots + p_{2N}X_n \to q_{21}X_1 + q_{22}X_2 + \cdots + q_{2k}X_k \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
\mathcal{R}_r &: p_{r1}X_1 + p_{r2}X_2 + \cdots + p_{rk}X_n \to q_{r1}X_1 + q_{r2}X_2 + \cdots + q_{rk}X_N
\end{aligned}$$

where the coefficients p_{ij} correspond to the number of species consumed in the reaction and the q_{ij} correspond to the number of species created in the reaction, and each reaction has a kinetic constant k_i associated to the reaction. By mass-action we mean that the rate of a chemical reaction is proportional to the product of the reactant numbers raised to the number of species created or consumed. We consider $X_t = (X_{1t} \ldots X_{kt})^T$ as a vector of the counts of the k species at time t. One can define a k by r matrix $S = (P - Q)^T$ with $P_{ij} = p_{ij}$ and $Q_{ij} = q_{ij}$. Given the stoichiometry matrix S one can either model the discrete problem $\Delta X = S\Delta R$ where ΔR is the vector of reaction events in a time step or the continuos problem $\mathrm{d}X_t = S\,\mathrm{d}R_t$. In either the discrete or continuous problem, a sequence of observations $\{Y_1, \ldots, Y_T\}$ either with or without noise are generated, and a common objective is to infer the kinetic constants as well as the matrices P and Q. This example is of interest because variants of this model have been used for continuous and discrete time dynamical systems with and without noise [7, 170, 172].

3. No noise

If no noise is present in the model (2.1)–(2.2), then we are in the setting of dynamical systems and ergodic theory (for general references, see [20, 87, 131, 168]). We first discuss the discrete-time setting and defer the material specific to the continuous-time setting until Section 4.5. Although such systems are deterministic in nature, there remains uncertainty regarding the initial condition of the system and also regarding which system (or parameter) is controling the observations. The long-range dependence inherent in such models is certainly not appropriate to all settings, but it is nonetheless realistic for many physical systems [3, 86, 110, 115, 128, 129, 139, 159, 178].

We will assume that the process $(X_n, Y_n)_n$ may be written with the following state-space formalism:

$$Y_n = f_a(X_n) \tag{3.1}$$

$$X_{n+1} = T_a(X_n), \tag{3.2}$$

where $T : \mathcal{A} \times \mathcal{X} \to \mathcal{X}$ is a parametrized family of maps and $f : \mathcal{A} \times \mathcal{X} \to \mathcal{Y}$ is a parametrized family of observation functions. Note that the model (3.1)–(3.2) is one of the classical objects of study in dynamical systems and ergodic theory.

3.1. Non-parametric system reconstruction from direct observations

Here we consider non-parametric estimation of a map T from direct observation of a single trajectory. That is, we observe a sequence $x_0, \ldots, x_n$, with $x_k = T^k(x_0)$, and we would like to estimate the map T in some sense. Although the methods discussed in this section do not directly involve parameter estimation, they are nonetheless relevant for parameter estimation, since any non-parametric method for estimation of a map immediately yields a method of parameter estimation if the map to be estimated comes from a parametrized family.

Let us first consider a case when the system can be successfully reconstructed from observations. If $\mathcal{X}$ is a manifold, T is continuous, the trajectory $(x_n)_n$ is dense in $\mathcal{X}$, and we observe the trajectory directly (*i.e.* the observations $(y_n)_n$ satisfy $x_n = y_n$), then T can be consistently estimated from $(y_n)_n$ using locally linear functions of the data. More precisely, let us state a result from [4] justifying this statement in the case $\mathcal{X} = [0,1]$. Let λ be Lebesgue measure on $[0,1]$. The map $T : [0,1] \to [0,1]$ is said to be an $E\{I_j, \alpha_j\}$-map if there exists at most countably many disjoint open intervals I_j and real numbers α_j such that $\lambda(\cup I_j) = 1$ and $f'(x) = \alpha_j$ for all x in I_j.

Proposition 3.1 ([4]). *Let T be an $E\{I_j, \alpha_j\}$-map. Suppose the observed trajectory $(x_n)_n$ is dense in $[0,1]$. Then there exists a sequence of estimates $\hat{T}_n$ of T such that for almost every x in $[0,1]$, it holds that $\hat{T}_n(x) = T(x)$ for all but finitely many n. In particular, $\hat{T}_n$ converges to T pointwise almost everywhere, and $\lambda(\{x : \hat{T}_n \neq T(x)\})$ tends to zero.*

To get an idea about how to prove this proposition, notice that for any two consecutive points x_n and x_{n+1} in the trajectory, the pair (x_n, x_{n+1}) lies on the graph of T. Therefore one may estimate T by linearly interpolating between neighboring points on the graph of T.

When the map T is not assumed to be continuous but only measurable, estimation of T from discrete observations of a single trajectory has been carried out by Adams and Nobel [4]. Their main result may be stated as follows.

Theorem 3.2 ([4]). *Suppose the system $(X, \mathcal{F}, \mu, T)$ is ergodic. Let μ_0 be a reference probability measure on $\mathcal{X}$ that is assumed to be "known." Also assume that there is a "known" constant M such that $1/M \leq d\mu/d\mu_0 \leq M$. Let $\mathrm{Meas}(\mathcal{X})$ denote the space of measurable functions from $\mathcal{X}$ to $\mathcal{X}$. Then there is an estimation scheme $(T_n)_n$ (whose definition uses M and μ_0), where $T_n : \mathcal{X}^n \to \mathrm{Meas}(\mathcal{X})$, such that for μ_0-a.e. initial condition x_0, the map $T_n(x_0, \ldots, x_{n-1})$ converges to T in a weak topology (i.e. $\mu\big(T_n^{-1}(A) \triangle T^{-1}(A)\big)$ tends to zero as n tends to infinity for each Borel set A).*

The estimation scheme $(T_n)_n$ that appears in [4] is constructed using an adaptive histogram method, which we discuss below. This paper also shows that under the same hypotheses the conclusion of the theorem is false if one requires that $\mu(\{x \in \mathcal{X} : T_n(x) \neq T(x)\})$ tends to zero as n tends to infinity.

Here we give an idea of the estimation scheme used in the proof of Theorem 3.2. The histogram method described here is actually from [122], which is very

similar in spirit to the method used in the proof of Theorem 3.2. Assume that $\mathcal{X} \subset \mathbb{R}^d$, and we fix a refining sequence $(\pi_k)_k$ of finite partitions of $\mathcal{X}$ with some additional properties (see [4] for details). Let $\pi_k(x)$ denote the cell in π_k containing x. Given the first n terms of the trajectory $(x_j)_{j=0}^{n-1}$, let

$$\phi_{n,k}(x) = \frac{\sum_{j=0}^{n-1} x_{j+1} I_{\{x_j \in \pi_k(x)\}}}{\sum_{j=0}^{n-1} I_{\{x_j \in \pi_k(x)\}}},$$

where $I_{\{x_j \in \pi_k(x)\}}$ is the indicator function of the event that x_j is in $\pi_k(x)$, and if the cell $\pi_k(x)$ contains no points x_j, then $\phi_{n,k}(x) = 0$. Now consider the empirical loss of $\phi_{n,k}$:

$$\Delta_{n,k} = \left(\frac{1}{n} \sum_{j=0}^{n-2} (\phi_{n,k}(x_j) - x_{j+1})^2 \right)^{1/2}.$$

The estimates $\hat{T}_n$ of T are adaptively chosen from among the $\phi_{n,k}$ according to $\Delta_{n,k}$ (using μ_0 and M). This method has the advantage that it works in quite a general setting (the only assumptions involve ergodicity and the Radon-Nikodym derivative with respect to a reference measure). On the other hand, it relies on the ergodic theorem for convergence, and therefore it appears very unlikely that it would have any general speed of convergence.

3.2. Non-parametric system reconstruction from general observations

In this section we consider approaches to system reconstruction when the observations $(y_n)_n$ are not necessarily equal to the trajectory $(x_n)_n$. There is a vast amount of literature on the technique of system reconstruction via delay coordinate embeddings. These system reconstructions may be thought of as non-parametric inference of dynamical systems. Delay coordinate embeddings are a well-studied inference procedure to reconstruct dynamical systems that satisfy certain conditions. In this section we define delay coordinate embeddings, mention some of the main uses of these techniques, and provide some representative theorems that provide conditions under which these methods work. Note that these reconstructions can help with the problems of feature estimation and possibly prediction, but they are not designed to help with parameter estimation directly.

The eventual goal of delay coordinate embedding techniques is typically feature estimation, which we summarize as follows. If the underlying map T and the observation function are both smooth, then under generic conditions, a delay coordinate embedding allows one to construct a smooth map $\widetilde{T}$ such that $\widetilde{T}$ is related to T by a smooth change of coordinates. Under this scenario, T and $\widetilde{T}$ will share many features, including entropy, Lyapunov exponents, and fractal dimensions of corresponding invariant measures. As these features are considered important in many physical settings, such delay coordinate reconstructions have been extensively studied.

To be specific, we consider a smooth map $T : \mathcal{X} \to \mathcal{X}$ of a manifold $\mathcal{X}$, with a smooth observation function $f : \mathcal{X} \to \mathbb{R}$. The data are assumed to be generated as follows: there is a trajectory $(x_n)_n$ such that $x_{n+1} = T(x_n)$, and we observe the data $(y_n)_n$ such that $y_n = f(x_n)$. The original idea to use delay coordinate embeddings to construct a system equivalent to $(\mathcal{X}, T)$ from the observations is due to Ruelle, at least according to the influential paper [128].

Definition 3.3. A delay coordinate mapping of $\mathcal{X}$ into $\mathbb{R}^m$ is a mapping $F : \mathcal{X} \to \mathbb{R}^m$ such that

$$F(x) = (f(x), f \circ T^{\tau}(x), \ldots, f \circ T^{\tau(m-1)}(x)),$$

for some natural number τ. The mapping F is said to be an embedding if it is a diffeomorphism from $\mathcal{X}$ to its image $F(\mathcal{X})$, that is if F is a smooth injection and has a smooth inverse.

The well-known theorem of Takens [159] (often called the Takens Embedding Theorem) may be stated as follows.

Theorem 3.4 ([159])**.** *If T, f, and τ satisfy certain genericity conditions and $m > 2\dim(\mathcal{X})$, then F is an embedding.*

Let $\tilde{\mathcal{X}} = F(\mathcal{X})$ and $\tilde{T} = F \circ T \circ F^{-1}$. The fact that F is an embedding means that the system $(\mathcal{X}, T)$ is related to the system $(\tilde{\mathcal{X}}, \tilde{T})$ by a smooth change of coordinates (given by F). In particular, invariants of $(\mathcal{X}, T)$ that depend on the differential structure of T (such as Lyapunov exponents or fractal dimensions of attractors) are equal to those of the system $(\tilde{\mathcal{X}}, \tilde{T})$.

This method of extracting invariants is typically carried out as follows. Given the data $(y_k)_{k=0}^{n-1}$, we may build time series data $(s_k)_{k=0}^{n-1-\tau(m-1)}$ for the system $(\tilde{\mathcal{X}}, \tilde{T})$ as follows: for $k = 0, \ldots, n-1-\tau(m-1)$, let

$$s_k = (y_k, y_{k+\tau}, \ldots, y_{k+\tau(m-1)}).$$

Then the new time series $(s_k)_k$ may be used to estimate invariant features of $(\tilde{\mathcal{X}}, \tilde{T})$, which will be the same as those features of $(\mathcal{X}, Q)$.

Takens' theorem has been generalized in various directions, such as filtered delay embeddings (see [149], for example) or delay embeddings for stochastic systems (see [155]), but we do not attempt to record all such results. However, the following generalization, due to Sauer, Yorke, and Casdagli, bears mentioning.

Theorem 3.5 ([149])**.** *Let A be a compact subset of $\mathcal{X}$ with box-counting dimension d. Let $m > 2d$. Suppose T, f, τ, and A satisfy certain genericity conditions. Then the delay coordinate map F given above is an injection on A and an immersion on each compact subset of any smooth manifold contained in A.*

The advantage of this theorem over the Takens theorem is that the relevant dimension d might be less than the ambient dimension of $\mathcal{X}$, in which case the

number of coordinates m required in the embedding space may be less than the number of coordinates required by Takens's theorem.

In order to use the delay coordinate method given only the data $(y_k)_{k=0}^{n-1}$, one must choose an appropriate dimension m and an appropriate lag τ. A variety of statistical techniques have been proposed to estimate the dimension m and find a suitable lag τ (for example, see the book [86] or the collection [115]), but further pursuit of these topics lies outside the scope of this survey.

3.3. Results from ergodic theory

In this section, we state some results from ergodic theory that are potentially relevant for parameter inference.

One of the most general results in this area is due to Ornstein and Weiss [126]. In this work, the authors consider the problem of estimation of stationary ergodic processes. (Note that in the setting of (3.1)–(3.2), if X_0 is distributed according to an ergodic invariant measure for T_a, then the observation process $(Y_n)_n$ satisfies exactly these conditions, as in Remark 2.9.) To make this problem precise, they consider the $\overline{d}$ metric on the space of such processes (see [126] for the definition of the $\overline{d}$ metric). Their main results may be stated as follows. First, they construct a procedure which, given a realization $(X_k)_{k=0}^{n-1}$ of a process $(X_k)_k$, constructs a process $Z^n = (Z_k^n)_k$. Then they show that the sequence of processes $(Z^n)_n$ converges to $(X_k)_k$ in the $\overline{d}$ metric if and only if $(X_k)_k$ lies in a certain class of processes, called Bernoulli processes. Thus, they have shown that there is a consistent estimation procedure for the class of Bernoulli processes. Furthermore, they show that no estimation procedure can be consistent for the class of all stationary ergodic processes.

In another direction, Ornstein and Weiss [125] show that entropy is the only finitely observable invariant in the following sense. Let J be a function from the class of finite-valued stationary ergodic processes to a complete separable metric space such that J is constant on isomorphism classes. The main result of [125] states that if J is finitely observable, then it must be a continuous function of the entropy. This result shows that there are strong restrictions on the possibilities for inference of isomorphism invariants.

Gutman and Hochman [63] extend the results in [125] in several ways. They give several rich families of classes $\mathcal{C}$ of stationary ergodic processes such that if J is a finitely observable invariant on $\mathcal{C}$, then J is constant. They also show that for every finitely observable invariant J on the class of irrational circle rotations, J is constant on the processes arising from a full measure set of angles. In particular, there is no finitely observable invariant for irrational rotations which is complete.

There is a large body of work, often categorized as smooth ergodic theory (see, for example, [9]), that seeks to understand the statistical properties of smooth (or piecewise smooth) dynamical systems. The typical setting is that one has a compact Riemannian manifold M and a smooth self-map $f : M \to M$. The manifold typically has a distinguished probability measure λ, which one may think of as volume measure on the manifold. The goal is to understand the

asymptotic behavior of the trajectory $\{f^n(x)\}_n$ for λ-a.e. x. For a wide class of such systems [176], often called (non-uniformly) hyperbolic systems, there is an invariant measure μ on M such that for x in a set of positive λ-measure, the trajectories in x equidistribute with respect to μ. In such cases, the measure μ is said to be a *physical* measure. Often the measure μ has some additional properties (it has no zero Lyapunov exponents and absolutely continuous conditional measures with respect to λ on unstable manifolds), and in this case μ may be called an SRB (Sinai-Ruelle-Bowen) measure [177]. The ergodic theory of SRB measures is fairly well-studied, and many of their statistical properties have been analyzed.

Example 3.6. The horseshoe map provides a prototypical example of a smooth, hyperbolic dynamical system. To describe the map, one begins with a topological disk, as seen on the left in Figure 2. One then shrinks the disk vertically by some factor $\lambda_v < 1/2$, expands the disk horizontally by a factor $\lambda_h > 2$, and wraps the resulting set back around inside the original disk as shown on the right in Figure 2. The set of points with interesting dynamics is the set of points that remain in the central square inside $\mathcal{X}$ for all time, which we denote by $\mathcal{C}$. Note that $\mathcal{C}$ is a topological Cantor set. Furthermore, observe that if x is in $\mathcal{C}$, then locally around x, the vertical direction is contracted, whereas the horizontal direction is expanded. Differentiable systems with this property (the tangent space can be divided into contracting and expanding directions) are generally called hyperbolic, and such systems tend to produce chaotic behavior.

FIG 2. *Depiction of a horseshoe map. One begins with a topological disk, as seen on the left. One then shrinks the disk vertically by some factor $\lambda_v < 1/2$, expands the disk horizontally by a factor $\lambda_h > 2$, and wraps the resulting set back around inside the original disk as shown on the right. Here we have chosen $\lambda_v = \frac{1}{5}$ and $\lambda_h = \frac{11}{5}$*

A related topic that has received a great deal of attention recently is that of statistical properties of dynamical systems (see, for example, [33]), especially concentration inequalities for dynamical systems [27, 28, 29, 30, 31]. These inequalities are used to study the fluctuations of observables for dynamical systems and have been shown to hold for sufficiently regular observables and a wide class of non-uniformly hyperbolic dynamical systems. Using these inequalities, it is possible to perform statistical estimation of various features of the dynamical system. See the survey [27] for more details and precise statements.

3.4. Parameter inference via synchronization and control

Synchronization-based approaches to parameter estimation have appeared quite often in the physics and control systems literature, *cf.* [3, 110, 129, 139, 178]

and references therein. In situations when these methods are used, it is common that no particular noise model is assumed. Indeed synchronization-based approaches are typically described as parameter inference methods in the noiseless setting, although they may be applied in other settings. The main idea of synchronization-based methods is to insert a "control" term in the defining equations of the system that allows one to incorporate the data. The parameter estimation may then be framed as a large optimization procedure in which one tries to find trajectories of the system which are close to the data. We do not know of rigorous theoretical justifications for this approach.

The topic of parameter estimation in a noiseless setting is discussed directly in the work of Abarbanel, Creveling, Farsian, and Kostuk [2], and we review their approach in this section. The main issue in this context is that one only has access to the observations $(Y_n)_n$, which might "hide" some information about the underlying system. The approach taken in [2] involves synchronization of the observations and the output of a model over the relevant time window. This approach may be summarized as follows.

Suppose that $\mathcal{X}$ is in $\mathbb{R}^d$ and the system (3.1)–(3.2) has the following form:

$$\begin{aligned} Y_n &= X_{n,1} \\ X_{n+1,i} &= T_{a,i}(X_n), \end{aligned}$$

where $X_{n,i}$ denotes the i-th coordinate of X_n. The synchronization approach taken in [2] is to add a "control" term of the form $k(Y_n - X_{n,1})$ to first coordinate of the model as follows:

$$\begin{aligned} \tilde{X}_{n+1,1} &= T_{a,1}(\tilde{X}_n) + k(Y_n - \tilde{X}_{n,1}) \\ \tilde{X}_{n+1,i} &= T_{a,i}(\tilde{X}_n), \quad i > 1. \end{aligned}$$

For $k > 0$ large enough, the data Y_n and the first coordinate $\tilde{X}_{n,1}$ of the model trajectory will "synchronize." With a fixed k, the authors propose to estimate the parameter a and the initial state X_0 by minimizing the following function:

$$C(a, X_0) = \sum_{j=0}^{n-1} (Y_n - \tilde{X}_{n,1})^2,$$

where the trajectory $\tilde{X}_n$ is computed starting at $\tilde{X}_0 = X_0$. The purpose of adding the control term is to regularize the function C so that its minimum may be found efficiently. Of course, the trajectory $\tilde{X}_n$ associated with this minimum is not a true trajectory of the original system. Therefore the authors propose a synchronization method that allows the parameter k to depend on time. In other words, they propose to minimize the cost function

$$C(a, X_0) = \sum_{j=0}^{n-1} (Y_j - \tilde{X}_{j,1})^2 + k_j^2,$$

subject to the constraints

$$\tilde{X}_{j+1,1} = T_{a,1}(\tilde{X}_j) + k_j(Y_j - \tilde{X}_{j,1})$$

$$\tilde{X}_{j+1,i} = T_{a,i}(\tilde{X}_j), \quad i > 1.$$

Variations of this method has been observed to work sufficiently well in practice. For example in [2], they observed good performance of more sophisticated versions of this method on a chaotic Colpitts oscillator and on a Hodgkin-Huxley neuron model. We also remark that to the best of our knowledge there are no theoretical guarantees regarding the consistency or performance of this method.

4. Observational noise only

If only observational noise is present in the model (2.1)–(2.2), then the underlying (deterministic) dynamical system still plays a critical role in determining the statistical properties of the system. We consider the following state-space formulation:

$$Y_n = f_a(X_n, \epsilon_n) \tag{4.1}$$

$$X_{n+1} = T_a(X_n), \tag{4.2}$$

where $(\epsilon_n)_n$ is a noise process, $T : \mathcal{A} \times \mathcal{X} \to \mathcal{X}$ is a parametrized family of maps, and $f : \mathcal{A} \times \mathcal{X} \times \mathcal{N} \to \mathcal{Y}$ is a parametrized family of noisy observation functions. Multiple authors explicitly argue for consideration of the observational noise model. For example, Judd [75] states that "the reality is that many physical systems are indistinguishable from deterministic systems, there is no apparent small dynamic noise, and what is often attributed as such is in fact model error." Furthermore, Lalley and Nobel [101] remark that "estimation in the observational noise model has not been broadly addressed by statisticians, though the model captures important features of many experimental situations." Additionally, applied works that take this point of view include [37, 58, 92, 94, 105, 106, 133, 148].

A distinguishing feature of the observational noise model is that the process $(X_n)_n$ is deterministic, and therefore in general it exhibits a long-range dependence structure. Furthermore, this long-range dependence is still present beneath the noise in the observation process $(Y_n)_n$. Such dependencies imply that traditional statistical estimation techniques do not apply and might not work. As Lalley and Nobel state in [101], "though some features of denoising can be found in more traditional statistical problems such as errors in variables regression, deconvolution, and measurement error modeling (*cf.* [23]), other features distinguish it from these problems and require new methods of analysis." In particular, they cite the facts that the covariates X_n are deterministically related (as opposed to i.i.d. or mixing), the noise is often bounded (as opposed to Gaussian), and the noise distribution itself is often unknown.

Example 4.1. Let $\mathcal{X} = [0, 1]$, and let $T_a : \mathcal{X} \to \mathcal{X}$ be given by $T_a(x) = ax(1-x)$, with a in $\mathcal{A} = [0, 4]$. This family of maps, known as the *logistic family*, has been extensively studied in a variety of settings. For $a \in [0, 1]$, it is known that for all x in $[0, 1]$, the iterates $T_a^n(x)$ tend to 0 as n tends to infinity. We say that a parameter value a has an attracting periodic orbit $\{p_0, \ldots, p_{N-1}\}$

if $T_a(p_i) = p_{i+1}$ (with indices interpreted modulo N) and $|(T_a^N)'(p_i)| < 1$. For such parameter values, the iterates $T_a^n(x_0)$ of Lebesgue almost every initial point x_0 will tend to the periodic orbit $\{p_0, \ldots, p_{N-1}\}$ as n tends to infinity. It is known [61, 103] that the set of parameter values that have an attracting periodic orbit is open and dense in $[0, 4]$. On the other hand, there are parameter values that give rise to very different asymptotic dynamics. In particular, we say that a parameter value a has an absolutely continuous invariant measure (acim) μ_a if μ_a is absolutely continuous with respect to Lebesgue and μ_a is an invariant measure for T_a. In such cases, it can be shown that the iterates $T_a^n(x_0)$ of Lebesgue almost every initial point x_0 equidistribute with respect to μ_a. Intuitively, the presence of μ_a produces seemingly stochastic behavior, which is often referred to as chaos. Jakobson showed in [71] that the set of parameter values that have an acim has positive measure in $[0, 4]$, and Lyubich later showed in [104] that Lebesgue almost every parameter in $[0, 4]$ either has an attracting periodic orbit or an acim.

In most of the papers cited in this section, this family of maps is taken as a standard testing ground for parameter estimation methods. Generally, it is assumed that the observational noise is additive (*i.e.* $f_a(x, \epsilon) = x + \sigma(a)\epsilon$).

4.1. Noise reduction

One basic approach to parameter estimation in the observational noise case is to reduce the noise and then apply parameter estimation methods. If the noise can be uniformly and sufficiently reduced, then these approaches will be approximately as successful as the estimation method applied to the noiseless case (but recall that there is very little statistical theory in the noiseless setting). For example, the positive results in [99, 100, 101] might be combined with a parameter estimation method in order to produce consistent estimates. Among the results contained in these works, the main positive result of [101] is the most general, and we state it as follows.

A homeomorphism F of a compact metric space (Λ, d) is said to be expansive with separation threshold Δ if for every $x \neq y$ in Λ, there exists n in $\mathbb{Z}$ such that $d(F^n(x), F^n(y)) > \Delta$. In the work [101], the authors consider an initial condition x and let $x_i = F^i(x)$. Also, they define a particular denoising algorithm, which, given observations $(y_i)_{i=0}^{n-1}$ obtained under an additive noise model, produces estimates $\hat{x}_{i,n}$ of the true states x_i. In this context, the main positive result may be stated as the following theorem.

Theorem 4.2 ([101])**.** *Let $F : \Lambda \to \Lambda$ be an expansive homeomorphism with separation threshold $\Delta > 0$. Suppose that the noise process $(\epsilon_n)_n$ satisfies $|\epsilon_n| \leq \Delta/5$ for every n. If $k = k(n) \to \infty$ and $k/\log(n) \to 0$ as n tends to infinity, then*

$$\frac{1}{n-2k} \sum_{i=k}^{n-k} |\hat{x}_{i,n} - x_i| \to 0, \qquad \text{as } n \to \infty$$

with probability 1 *for almost every initial point* x *in* Λ *(with respect to any* F *invariant Borel probability measure).*

By allowing a slight modification to their estimation scheme, the authors also show that under the same hypotheses

$$\max_{\log(n)\leq i\leq n-\log(n)} |\hat{x}_{i,n} - x_i| \to 0, \qquad \text{as } n \to \infty$$

with probability 1 for almost every initial point x in Λ.

Of course, the task of removing the noise might itself be difficult or in some cases even impossible, as witnessed by the negative results in [99, 100, 101] and the related results in [76, 77, 79, 80]. Here we state the main negative result in [101]. A pair of points x and x' is said to be strongly homoclinic for the homeomorphism F if

$$\sum_{k\in\mathbb{Z}} d(F^k(x), F^k(x')) < \infty.$$

Theorem 4.3 ([101])**.** *Suppose the stationary distribution for the noise process* $(\epsilon_n)_n$ *is unbounded (or has sufficiently large support). If* x *and* x' *are strongly homoclinic, then for every measurable function* $\phi : \prod_n \mathcal{X} \to \mathbb{R}^d$,

$$\mathbb{E}\Big[|\phi((y_n)_n) - x| - |\phi((y'_n)_n) - x'|\Big] > 0.$$

In other words, even with access to the entire observation sequence, any state estimation or denoising scheme will fail with positive probability.

In addition to the works mentioned so far in this section, the following works discuss methods and numerical simulations related to the problem of denoising or smoothing data in the presence of only observational noise: [37, 58, 92, 94, 105, 106, 148].

4.2. Introduction to likelihoods and related methods

We begin with the work of Berliner [10, 11], which sets the stage for most of the work that has followed. In [10], the author is mostly concerned with the observational noise setting (4.1)–(4.2). The likelihood function is given by

$$L(x_0, a) = p(y_0^{n-1}|x_0, a),$$

where $p(y_0^{n-1}|x_0, a)$ denotes the likelihood of observing y_0^{n-1} given the parameter choice a and the true initial condition x_0 (*i.e.* $p(\cdot|x_0, a)$ is the probability density for the observation process conditional on x_0 and a). Depending on the context, there may be different parameter estimation methods that go by the name maximum likelihood estimation. Some authors refer to the maximum likelihood (ML) method for estimating the parameter a when considering the following maximum likelihood estimator (MLE):

$$\hat{a}_n = \operatorname*{argmax}_a \max_{x_0} L(x_0, a).$$

On the other hand, if the parameter a also corresponds to an initial distribution π_a for x_0, as in [111], then one may refer to the ML method when considering the following marginalized MLE:

$$\hat{a}_n = \operatorname*{argmax}_a \int L(x_0, a)\, d\pi_a(x_0).$$

It will be useful to find an explicit form for the likelihood function in the case that (i) the observational noise sequence $(\epsilon_n)_n$ is assumed to be i.i.d. normal with zero mean and unit variance, and (ii) the observation function f_a takes the form $f_a(x, \epsilon) = x + \sigma(a)\epsilon$. The function $\sigma(a)$ allows one to set the variance of the noise according to the parameter a. In this case, we have

$$L(x_0, a) = \left(\sigma(a)\sqrt{2\pi}\right)^{-n} \exp\left(-\sum_{k=0}^{n-1}(y_k - T_a^k(x_0))^2/(2\sigma^2(a))\right)$$

and the corresponding log-likelihood function is given by

$$\log L(x_0, a) = -n \log\left(\sigma(a)\sqrt{2\pi}\right) - \sum_{k=0}^{n-1}(y_k - T_a^k(x_0))^2/(2\sigma^2(a)).$$

A significant portion of the work on parameter estimation following Berliner has involved optimization of this log likelihood function, even when the noise is not necessarily Gaussian and thus its interpretation as a log likelihood function is no longer valid.

As discussed in [133], standard statistical results do not apply to the ML method in this setting. With the above notation, the main difficulty in the current setting is that T_a^k is a non-stationary function of k. Standard statistical results on the performance of the ML method apply when the likelihood function has no such dependence on k (or is periodic with respect to k), but these results do not apply *a priori* in the current setting.

Under suitable conditions on the dynamical systems and the observations, it has recently been shown that (marginalized) maximum likelihood parameter estimation is consistent [111]. The proof involves ideas from both information theory and dynamical systems. Furthermore, in the same work, the authors show how some well-studied properties of dynamical systems imply certain general statistical properties related to maximum likelihood estimation. Lastly, the authors exhibit classical families of dynamical systems for which maximum likelihood estimation is consistent. Examples include shifts of finite type with Gibbs measures and Axiom A attractors with SRB measures.

The Bayesian approach assumes a prior distribution (density) for x_0 and a, written as $\pi(x_0, a)$. Given the data y_0^{n-1}, the posterior distribution is then

$$\pi(x_0, a|y_0^{n-1}) = \frac{p(y_0^{n-1}|x_0, a)\pi(x_0, a)}{\int p(y_0^{n-1}|x, a)\pi(x, a)\, dxda}.$$

With these basic definitions, Berliner considers three main methods of parameter estimation: maximum likelihood estimation, minimization of a cost function (which is often chosen to be the negative of the log likelihood function)

and Bayesian estimation. One of Berliner's main points is that when the system (4.1)–(4.2) is chaotic, the likelihood function will also typically be chaotic, in the sense that it will be extremely jagged. The rough nature of these likelihood functions makes all three of the above methods of statistical estimation computationally very expensive, and much of the work following Berliner has been motivated by the need to mitigate this difficulty. Beyond these computational difficulties, we remark that to our knowledge the only general theoretical results concerning the consistency of any of these likelihood-based methods appears in [111].

4.3. Variations on likelihood based methods

A common method of parameter estimation in practice is to minimize some cost function C with respect to the parameters. Given the observations $(y_k)_{k=0}^{n-1}$, such methods employ the following estimators:

$$\hat{a}_n = \operatorname{argmin}_a \min_{x_0} C\left(x_0, a, (y_k)_{k=0}^{n-1}\right),$$

where $C(x_0, a, (y_k)_{k=0}^{n-1})$ somehow measures the discrepancy of the observations and the system trajectory having parameter a and initial state x_0.

As we mentioned in the previous section, the most basic cost function is the least squares cost function

$$C_{LS}\left(x_0, a, (y_k)_{k=0}^{n-1}\right) = \sum_{k=0}^{n-1} (y_k - T_a^k(x_0))^2.$$

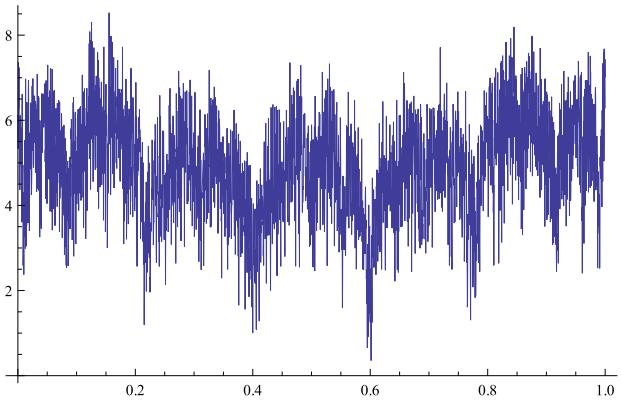

FIG 3. *Least Squares cost function for x_0 in logistic family as a function of $x \in [0,1]$ given $n = 20$ observations, true initial value $x_0 = .4$ and true parameter $a = 4$*

Due to the form of the cost function C_{LS}, it might appear that the theory of non-linear least squares (*cf.* [135] and references therein) could be used to prove

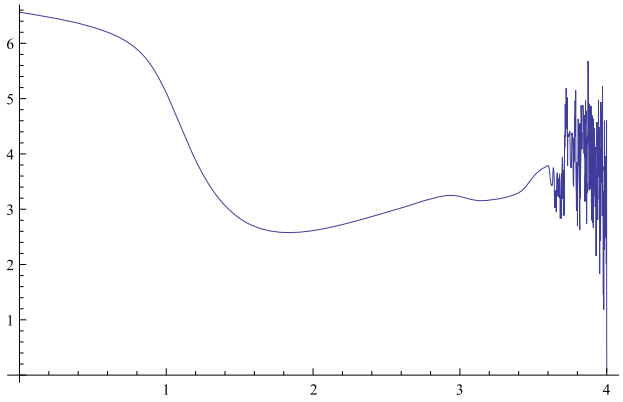

FIG 4. *Least Squares cost function for parameter a in logistic family as a function of $a \in [0,4]$ given $n = 20$ observations, true initial value $x_0 = .4$ and true parameter $a = 4$*

consistency of the estimator that minimizes C_{LS}. Unfortunately, the existing results in this theory do not typically yield meaningful results when applied to the sequence of maps $\{T_a^k\}_k$. In particular, by the product rule, one typically expects that in hyperbolic dynamical systems the derivatives $\partial_a T_a^k(x)$ will grow exponentially as a function of k, and such growth implies that the estimates of [135] cannot be used even to show consistency. Therefore it remains an interesting open question as to whether this theory can be adapted to the dynamical systems setting.

Perhaps due to the sensitive dependence of C_{LS} on x_0 and the additional computational expense incurred by minimizing C_{LS} over x_0, several authors considered minimization of a one-step least squares cost function, given by

$$C_{OSLS}(a, (y_k)_{k=0}^{n-1}) = \sum_{k=0}^{n-2} (y_{k+1} - T_a(y_k))^2, \tag{4.3}$$

which does not depend on any initial condition x_0. This cost function may appear to be the familiar least squares function from regression analysis, but as Kostelich [93] recognized, it suffers from the problem of errors in variables (*cf.* [24, 57]). The problem of errors in variables is that the errors are not independent, as is implicitly assumed in the form of the cost function. Viewing C_{OSLS} from the perspective of traditional regression, we see that y_k appears to play the role of the independent variable and y_{k+1} plays the role of the dependent variable, but both y_k and y_{k+1} contain noise according to the model (4.1)–(4.2). It is well-known that the problem of errors in variables can lead to asymptotically biased results, and therefore we should not expect minimization of C_{OSLS} to give consistent estimates of the parameter a.

In response to the errors in variables problem, Jaeger and Kantz [70, 85] propose a "solution" of the problem, which amounts to minimizing the following

cost function that has since gone by the name "total least squares" cost function:

$$C_{TLS}(a,(y_k)_{k=0}^{n-1}) = \sum_{k=0}^{n-1} \min_{y\in\mathcal{X}} \|(y_k,y_{k+1}) - (y,T_a(y))\|^2.$$

Note that this approach essentially ignores the dynamics altogether, and instead focuses on minimizing the sum of orthogonal distances between the graph of T_a and the points (y_k,y_{k+1}) in $\mathcal{X}\times\mathcal{X}$. In order to include some aspect of the dynamics, they further modify their cost function to find local shadowing trajectories by considering cost functions of the form

$$C_{MTLS}(a) = \sum_{k=0}^{n-s-1} \min_{y} \|(y_k,\ldots,y_{k+s}) - (y,\ldots,T_a^s(y))\|^2.$$

Here s is a parameter of the method; it is the number of steps over which one considers the local shadowing trajectories. If one asks for global shadowing trajectories, corresponding to $s=n-2$, then this modified total least squares cost function is equivalent to the original least squares cost function C_{LS}.

McSharry and Smith [114] consider the one step cost function C_{OSLS} given by (4.3). They prove that in the case of the logistic map with a specific parameter value, the minimization of this cost function produces biased estimates, even with infinitely many observations. Their proposed solution involves minimizing the cost function given by

$$C_{MS}(a) = -\sum_{k=0}^{n-1} \log\left(\int \exp\left(-\frac{d_k^2(x)}{2\epsilon^2}\right)\mu_a(dx)\right),$$

where $d_k^2(x) = \|(y_k,y_{k+1}) - (x,T_a(x))\|^2$, ϵ is the variance of the noise process $(\epsilon_n)_n$, and μ_a is a particular invariant measure for the map T_a. They argue that the minimum of C_{MS} provides more reliable parameter estimates due to its inclusion of information regarding the invariant measure μ_a. It is perhaps a shortcoming of this method that one must know the variance of the noise process and the invariant measure μ_a in order to calculate $C_{MS}(a)$. In practice, the authors suggest approximating the integral with respect to μ_a by a sum over a long piece of trajectory simulated from the model in the hopes that this approximation will be close to the integral by the ergodic theorem. The authors provide numerical evidence that C_{MS} provides better parameter estimates than either C_{OSLS} or C_{TLS}, although again no theoretical results are available to justify this comparison.

Meyer and Christensen [116], following up on the work of McSharry and Smith [114], propose to model the system using a combined noise state-space model of the form (2.1)–(2.2), and proceed via an MCMC algorithm for performing the inference. In particular, they take a Bayesian approach, modeling both the true states X_n and the parameters a as unknown variables. They assume that the process $(X_n)_n$ forms a Markov chain (by adding dynamical noise to the model). Then they compute posterior probabilities of the unobserved variables

using the Gibbs sampler and the Metropolis-Hastings algorithm. This approach rasies the question of model misspecification: how do these estimators perform when the model does not accurately represent the noise structure (see Question 7.3)?

Several further works build on this stream of research, proposing related approaches for parameter estimation in the observational noise setting. For example, there are gradient descent methods [75, 78, 144], methods that involve cutting the time-series data into small subintervals and performing ML estimation on each interval independently [133], methods involving backwards iteration of the map [153], and iterative methods that alternate between estimating the system states and the system parameters [121].

4.4. Method of moments

Here we mention a method of parameter estimation that has been shown to be consistent at least for the logistic family, introduced in Example 4.1. For the observational noise model, this method, discussed in [133], provides a rare example of a method that has been proved to be consistent for at least one non-trivial example.

We consider the model (4.1)–(4.2), where $\mathcal{X} = [-1, 1]$, $\mathcal{A} = [0, 2]$, and $T_a(x) = 1 - ax^2$, which is re-parametrization of the family in Example 4.1. Assume that the underlying trajectory process $(X_n)_n$ is ergodic, which is the case if one assumes that X_0 is drawn from an ergodic invariant measure μ_a for the map T_a. Alternatively, one may assume that a is chosen such that T_a has an acim μ_a (as discussed in Example 4.1) and X_0 is drawn from Lebesgue measure. Also assume that the observational noise is additive (*i.e.* $Y_n = X_n + \epsilon_n$) and $(\epsilon_n)_n$ is i.i.d. Gaussian with mean 0 and variance ϵ^2. For a sequence $(z_k)_{k=0}^{n-1}$, let $A_n(z_k) = \frac{1}{n}\sum_{k=0}^{n-1} z_k$ and for any $f : \mathbb{R} \to \mathbb{R}$, let $\mathbb{E}_{\mu_a}(f) = \int f(x)\, d\mu_a(x)$. Then by the ergodic theorem

$$\lim_{n\to\infty} A_n(Y_k) = \mathbb{E}_{\mu_a}(x) \tag{4.4}$$

$$\lim_{n\to\infty} A_n(Y_k^2) = \mathbb{E}_{\mu_a}(x^2) + \epsilon^2 \tag{4.5}$$

$$\lim_{n\to\infty} A_n(Y_k^3) = \mathbb{E}_{\mu_a}(x^3) + 3\epsilon^2\mathbb{E}_{\mu_a}(x) \tag{4.6}$$

$$\lim_{n\to\infty} A_n(Y_kY_{k+1}) = \mathbb{E}_{\mu_a}(x) - a\mathbb{E}_{\mu_a}(x^3). \tag{4.7}$$

Also, averaging the equation $x_{n+1} = 1 - ax_n^2$, we obtain that

$$\mathbb{E}_{\mu_a}(x) = 1 - a\mathbb{E}_{\mu_a}(x^2). \tag{4.8}$$

Combining Equations (4.4)–(4.8), we arrive at the following estimates for the unknown parameters a, $\mathbb{E}_{\mu_a}(x)$, $\mathbb{E}_{\mu_a}(x^2)$, $\mathbb{E}_{\mu_a}(x^3)$, and ϵ:

$$\hat{a}_n = \frac{A_n(Y_kY_{k+1}) + 2A_n(Y_k) + 3\big(A_n(Y_k)\big)^2}{3A_n(Y_k)\big(A_n(Y_k)\big)^2 - A_n(Y_k^3)}$$

$$\begin{aligned}
\mathbb{E}_{\mu_a}(\hat{x})_n &= A_n(Y_k) \\
\mathbb{E}_{\mu_a}(\hat{x}^2)_n &= A_n(Y_k^2) - \hat{\epsilon}_n^2 \\
\mathbb{E}_{\mu_a}(\hat{x}^3)_n &= \frac{1}{\hat{a}_n}\big(A_n(Y_k) - A_n(Y_k Y_{k+1})\big) \\
\hat{\epsilon}_n &= \frac{A_n(Y_k^3) - \mathbb{E}_{\mu_a}(\hat{x}^3)_n}{3A_n(Y_k)}
\end{aligned}$$

These estimates are consistent by the ergodic theorem, but they might converge quite slowly, as there is no general rate of convergence in the ergodic theorem.

4.5. Ordinary differential equations with observational noise

In this section we consider systems such that the state of the system satisfies an ODE:

$$\dot{X}_t = F_a(X_t) \tag{4.9}$$
$$X_0 = x_0 \tag{4.10}$$

where, for simplicity, x_0 and X_t are in $\mathbb{R}^d$ and $F : \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}^d$ satisfies some regularity conditions depending on the context. Furthermore, one typically assumes that observations are of the form

$$Y_{t_k} = X_{t_k} + \epsilon_k,$$

where $0 \leq t_0 < \cdots < t_n$ and the measurement errors $(\epsilon_k)_k$ are i.i.d. with zero mean.

In contrast to the discrete-time setting, the topic of identifiability for ODE models has been widely studied. For a recent and comprehensive review of identifiability in ODEs, see [118]. Necessary and sufficient conditions for identifiability are known in the case that the system of equations is linear in the unknown parameters [36]. In many cases, especially in moderately large dimensions, structural identifiability may be difficult to establish. Thus, several recent works have investigated methods for parameter estimation (possibly of sets of parameters) in nonidentifiable settings [22, 97, 113, 143, 142, 154, 165].

To get an idea of the methods and difficulties involved, let us consider the problem of parameter estimation. With the above notation, the most straightforward procedure for estimation of the true parameter a_0 is the least-squares method, which is defined as

$$\hat{a}_n = \operatorname{argmin}_a \sum_{k=0}^{n} ||Y_{t_k} - x(t_k, a)||^2,$$

where $x(t, a)$ denotes the solution of Equations (4.9)–(4.10). In order to find these estimates in practice, one generally uses an optimization procedure that relies on numerical integration to find $x(t_k, a)$ for various choices of a. In [174],

the theory of non-linear least squares estimation is adapted to the current setting to show that under some regularity conditions, the (numerical) least-squares estimator is consistent and asymptotically normal. We note that they consider only "in fill" asymptotics for these results. Also, let us mention that despite the good properties mentioned above, the least-squares estimator may be very difficult or expensive to compute in practice.

Bayesian approaches have also been developed for parameter estimation in ODEs (see, for example, [59, 60]). In general, though, there are no analytical expressions for the posterior distribution, and therefore one must solve the (generally difficult) computational problem of numerically approximating potentially high-dimensional integrals with complex integrands.

Perhaps due to the heavy computational burden required to perform the numerical integration in the least-squares estimator or in the Bayesian approach, another type of estimator has been developed, which avoids numerical integration. These estimators are often referred to as regularization or collocation methods. Recent examples of such methods may be found in [62, 138, 140] (and references therein). We mention here some recent results, which appear in [62]. First, define the following estimates of X_t: for $t > 0$, let

$$\hat{x}(t) = \sum_{k=1}^{n} (t_k - t_{k-1}) \frac{1}{b} K\left(\frac{t - t_k}{b}\right) Y_{t_k},$$

where K is a suitable kernel and b is a bandwidth. Now define

$$\hat{a}_n = \text{argmin}_a \int_0^1 ||\hat{x}'(t) - F_a(\hat{x}(t))||^2 w(t) dt,$$

where $\hat{x}'(t)$ denotes the derivative of $\hat{x}(t)$ and w is a weight function. The main results of the paper show that under "in fill" asymptotics, this estimator is consistent with $\sqrt{n}$-rate. Despite these results, basic statistical properties of some of the earliest collocation methods (*e.g.*, [66, 166]) remain poorly understood. Furthermore, the rigorous results of [36, 62] do not address the statistical efficiency of their estimators, and such methods also typically require a choice of the smoothing/regularization parameter, which makes them non-trivial to apply to real data. Nonetheless, one of the advantages of such methods could be their ability to perform effectively even in the presence of small dynamical noise (see Question 7.3).

5. Dynamical noise only

In this section, we consider the case when only dynamical noise is present in the system. In particular, we consider the the following setting:

$$\begin{aligned} Y_n &= f_a(X_n) \\ X_{n+1} &= T_a(X_n, \delta_n), \end{aligned}$$

where $(\delta_n)_n$ is a noise process, $T : \mathcal{A} \times \mathcal{X} \times \mathcal{N} \to \mathcal{X}$ is a parametrized family of noisy maps, and $f : \mathcal{A} \times \mathcal{X} \to \mathcal{Y}$ is a parametrized family of observation functions. The dynamical noise model has been studied in the dynamical systems literature under the name "random dynamical systems" (see [90] and references therein). The process $(X_n)_n$ forms a discrete-time Markov chain on the continuous state space $\mathcal{X}$ (see the book of Meyn and Tweedie [117] and references therein). In this case, some of the estimation methods from the statistical literature on time series and state space models apply (see Section 6).

Without using this Markov structure, Adams and Nobel have studied the non-parametric reconstruction of such systems from direct observations (*i.e.* $Y_n = X_n$) [122, 123]. In particular, they used adaptive histogram methods to show results similar to those regarding non-parametric reconstructions of systems with no noise, as in Section 3.1. These methods do not work in the observational noise case precisely because in that setting they suffer from the problem of errors in variables, as discussed in Section 4.3.

Let us also mention a topic, called stochastic stability, that is often discussed in connection with random dynamical systems. A common setting for random dynamical systems is to assume that there is a map $T : \mathcal{X} \to \mathcal{X}$, where $\mathcal{X}$ is a compact manifold and T is smooth, with a "natural" invariant probability measure μ. In common examples, T might be a (non-uniformly) hyperbolic map and μ might have the property that almost every initial condition with respect a volume measure on the manifold equidistributes with respect to μ. In such cases, one typically adds dynamical noise as follows. Let $\epsilon > 0$. For each x in $\mathcal{X}$, let $\mathbb{P}_\epsilon(x, \cdot)$ be the uniform measure on the ball of radius ϵ about the point $T(x)$. Then the Markov chain corresponding to this random dynamical system is determined by viewing $\mathbb{P}_\epsilon$ as the transition kernel for the chain. Under some conditions, the chain corresponding to $\mathbb{P}_\epsilon$ will have a unique stationary distribution, μ_ϵ. A well-known result (see [90]) states that under certain conditions, the measure μ_ϵ converges to μ weakly as ϵ tends to 0, in which case the system is said to exhibit stochastic stability. To the best of our knowledge, no theoretical work on parameter estimation has been conducted for this particular setting, perhaps making it an area ripe for progress. On the other hand, this setting may be viewed as a particular case of the general state-space setting, in which there is no observational noise, and therefore methods described in Section 6 might also be applicable here.

6. General state space models

In this section we consider the full system (2.1)–(2.2), where both dynamical noise and observational noise are present. Specific versions of such models have long been considered in the statistics literature, where they are known as state space models [48]. The literature on state space models in both applied and theoretical statistics is extensive and [65, 134] are two excellent texts covering applied modeling on this topic. The models can be summarized as the study of hidden Markov models (HMMs) in general state-spaces. (For an article discussing the connections between ergodic theory and finite state HMMs, see [18].)

Theoretical understanding of general HMMs has been a challenge and rigorous statements on consistency in parameter estimation have only appeared recently [44] (see Section 6.2). Most of the work in this area has been devoted to the problem of state estimation or filtering, and even at a computational level the problem of parameter estimation is still largely unsolved. In this section we survey some of the most studied approaches to filtering and discuss parameter estimation where there are results.

6.1. Kalman filter and some generalizations

The simplest such models assume that the dynamics are linear and the noise is additive and Gaussian:

$$X_{n+1} = AX_n + B\delta_{n+1}$$
$$Y_n = CX_n + D\epsilon_n,$$

where here A, B, C, and D are all matrices of the appropriate dimension and $(\delta_n)_n$ and $(\epsilon)_n$ are independent i.i.d. Gaussian processes. In this case, the optimal solution to the state estimation or denoising problem is given by the well-known Kalman filter [50, 83]. Generalizations of the ideas behind Kalman filtering to non-parametric models have been an extensive area of research in Bayesian and frequentist inference [48, 55, 56, 119].

Conceptually, the simplest generalization of the Kalman filter to nonlinear models involves linearizing the models at each time point and then using the Kalman filter. This method is often called the *extended Kalman filter* (EKF) [72, 6]. While the Kalman filter is optimal in the sense that it is the minimal-variance unbiased estimator, the general EKF is known to be biased. Furthermore, due to the linearization of the model, the propagation of the error covariance estimates may behave quite poorly if the non-linear terms in the model are significant.

The unscented Kalman filter (UKF) [81, 82] provides a deterministic sampling scheme that has been observed to outperform the EKF. The basic idea behind the UKF is that instead of approximating the model by linearization, one ought to use the exact model but approximate the posterior distributions by Gaussian distributions. The sampling scheme is designed to insure that the first two moments of the posterior distributions match the first two moments of the approximating distributions. It is believed that the UKF outperforms the EKF because it may be viewed as an unbiased second-order method, whereas the EKF is a biased first-order method. Of course, the UKF is believed to have shortcomings of its own; in particular, it assumes that the posterior distributions are Gaussian, which is certainly not the case in general. Also, the number of samples required for the UKF is at least the dimension of the state space, and in high-dimensional settings this fact makes the UKF computationally intractable. A wide variety of Monte Carlo (MC) methods have been proposed to overcome these issues.

Another generalization of the Kalman filter is known as the ensemble Kalman filter (EnKF) [21, 51, 52]. This method is a Monte Carlo method that is partic-

ularly popular in the weather prediction community. In fact, this method may be thought of as a type of particle filter (see Section 6.4.1).

6.2. MLE for HMMs

If one is willing to consider point estimates of unknown parameters in a setting where the likelihood function is known, then one can consider the maximum likelihood method (MLE) for parameter estimation. Let us now state the main result of the paper [44], which gives sufficient conditions for the consistency of MLE in this context. Let $(X_k, Y_k)_{k=1}^{\infty}$ be a hidden Markov model (HMM) of the form (2.1)–(2.2). Let a^* denote a fixed parameter value in $\mathcal{A}$. Assume that the HMM with parameter a^* has a unique stationary distribution, and let $\mathbb{P}_{a^*}$ be the corresponding stationary HMM. Denote by $p^{\nu}(y_0^n, a)$ the likelihood of the observations Y_0^n with initial distribution $X_0 \sim \nu$ and parameter a. Consistency of the maximum likelihood estimator (MLE) may now be stated in the following form: if $a_n = \operatorname{argmax}_a p^{\nu}(y_0^n, a)$, then a_n converges $\mathbb{P}_{a^*}$-a.s. to a^* as n tends to infinity. The main result of [44] gives some general conditions under which the MLE is consistent in this sense. A precise statement of these general conditions is beyond the scope of this survey.

6.3. Bayesian inference

Recall the Bayesian formulation of state space estimation or filtering. Here one assumes that the model (2.1)–(2.2) gives rise to probability densities $\mu(x_0)$, $p(x|x')$, and $q(y|x)$, which define the initial distribution, transition kernel, and marginal distribution of the observation process, respectively. The densities are with respect to some fixed reference measures denoted dx and dy. In this framework, we are given access to finitely many observations y_0^{n-1}, and we would like to estimate the true trajectory x_0^{n-1}. Our assumptions define likelihood functions

$$p(x_0^{n-1}) = \mu(x_0) \prod_{k=0}^{n-2} p(x_{k+1} \mid x_k),$$

and

$$p(y_0^{n-1} \mid x_0^{n-1}) = \prod_{k=0}^{n-1} q(y_k \mid x_k).$$

Given the observations y_0^{n-1}, the posterior distribution for X_0^{n-1} is given by

$$p(x_0^{n-1} \mid y_0^{n-1}) = \frac{p(x_0^{n-1}, y_0^{n-1})}{p(y_0^{n-1})},$$

where

$$p(x_0^{n-1}, y_0^{n-1}) = p(x_0^{n-1})\, p(y_0^{n-1} \mid x_0^{n-1})$$

$$p(y_0^{n-1}) = \int p(x_0^{n-1}, y_0^{n-1}) dx_0^{n-1}.$$

There are a few instances when these distributions may be calculated analytically, such as when the system is linear and the noise is Gaussian or when $\{X_n\}_n$ is a finite state Markov chain. Outside of these cases, there is no analytical method for calculating the posterior distribution, and therefore one seeks a numerical approximation for this distribution. With the significant advances in computational power in recent years, there has been a remarkable amount of research devoted to finding efficient computational approaches to approximating such posterior distributions. Section 6.4 discusses some of the more recent computational approaches to filtering.

An interesting work in the Bayesian context is [150] where the author studies posterior consistency for dependent data from an information theoretic point of view. The author establishes posterior consistency for misspecified models under the assumption of asymptotic equipartition property. For finite state space ergodic models, this is implied by the Shannon-McMillan-Breiman theorem. It could be interesting and useful to extend the ideas from [150] to prove posterior consistency in parameter estimation for more general dynamical systems.

6.4. Inference for dynamical systems via simulation based methods

In the general non-linear, non-Gaussian state-space setting of (2.1)–(2.2), the posterior distributions for x_0^{n-1} are not available in closed form, as they involve some integrals for which no analytical evaluation methods exist. In order to perform inference in this setting, a great deal of effort has been devoted to developing sophisticated computational algorithms for sampling from these posterior distributions. One general idea is to use Monte Carlo (MC) methods to estimate the integrals of interest. It is worth emphasizing that there has been a huge amount of work in this direction, and we do not claim to provide a comprehensive survey of all the relevant results. For an introduction to MC methods, see the book [145].

6.4.1. MCMC methods, SMC and particle filters

If one cannot sample from the posterior distribution directly, then one often turns to Markov chain Monte Carlo (MCMC) methods. For a discussion of such methods, see the books [145, 169] and references therein. Such methods have been used for parameter estimation in dynamical systems (*e.g.*, [34]).

Traditional Monte Carlo or MCMC methods may be used to perform "batch" inference, *i.e.* when all of the observations are available at once and one would like to estimate $p(x_0^{n-1}|y_0^{n-1})$ for fixed n, although even in this setting they might be prohibitively computationally expensive. When the goal is to perform "on-line" or sequential inference, or in an effort to try to reduce the computational expense, one might try sequential Monte Carlo methods (SMC) and their

many variations. A particularly popular version of these methods is known as particle filtering. For a well-written, thorough introduction to the principles of sequential Monte Carlo (SMC) and particle filtering methods, see the recent tutorial by Doucet and Johansen [47]. For an incomplete list of works concerning SMC and particle filtering, as well as their adaptations to parameter estimation, see [21, 39, 41, 51, 52, 53, 68, 84, 102, 120, 124, 136, 157]. The basic idea is that the posterior distributions of interest are approximated by a finite collection of N samples, called particles, which are recursively propagated through the model. The main theoretical advantage of these methods is that one is often able to establish the convergence of the approximations to the true posterior distributions as the number of particles N tends to infinity.

6.4.2. ABC methods

Most of the methods mentioned previously in this section rely on explicit knowledge and evaluation of the likelihood function. In many situations, such as in high dimensional complex models, the likelihood function may not be available or is computationally expensive to evaluate. In such scenarios, a simple computational method called approximate Bayesian computation (ABC) offers a powerful alternative to conduct statistical inference. ABC was first proposed as a philosophical argument in [147] and introduced to population genetics in [160]. Since then these methods have become extremely popular in many applied fields. A partial list of references include [38, 40, 112, 130, 137, 141, 152, 163, 172]. A good review with applications to filtering is [109]. Briefly, in ABC methods one first draws a parameter value θ^* from the prior distribution and generates synthetic data from the likelihood model corresponding to θ^*. If the synthetic data "is similar to" the observed data (measured in some metric) up to a prespecified tolerance then θ^* is accepted as a draw from the (approximate) posterior distribution. Choosing the metric and the tolerance level are difficult problems, but partial results are known ([54]).

An important point to note is that in many examples, a summary statistic instead of the original data set is used for matching. This clearly results in loss of information (and sometimes even results in invalid inference; see [146]) and thus raises the interesting question about when one can perform consistent model selection using the ABC methodology. In [108] a sufficient criteria is worked out, but clearly more needs to be done especially in the context of dynamical systems.

6.5. Stochastic differential equations (SDE)

SDEs constitute an important class of modeling tool (see [12, 171, 67] and the references within) given by

$$dX_t = b(\theta, X_t)dt + \sigma(\theta, X_t)dW_t$$

where θ is an unknown parameter, b, σ are known functions and W_t is a Brownian motion on $\mathbb{R}^n$. The inference problem is to estimate θ from discrete observations of X_t observed with or without measurement error. We will be brief in our review for these class of models and point the reader to existing, comprehensive references and recent key papers. The theory for parameter estimation for such models is very mature for continuously observed X_t (where continuous observation means that one has access to X_t for all $t \in [0, T]$); see [17] for a book length treatment and an extensive list of references. The last decade witnessed a few breakthroughs for discretely observed X_t (where discretely observed means that one has access to X_t for discrete values $t \in \{t_0, \ldots, t_N\}$) [13, 12]. In [13] the authors develop an algorithm to "perfectly" sample the paths of X_t at any finite number of points, *i.e.*, sample the values of X_t at these points without any discretization error typically associated with numerical schemes such as the Euler scheme. Using this algorithm, the authors further develop a framework for statistical inference [12]. The approach taken in [12] is computationally intensive and might not scale well with dimension. See [95] for an alternative numerical approach which may be scalable.

7. Open questions and future directions

Here we list some open questions related to parameter inference in dynamical systems and discuss possible future research directions. In general, it seems that most existing statistical inference methods make use of the presence of dynamical noise in the system.

The first question concerns the topic of identifiability in discrete time. Recall that in contrast to the situation for discrete time, identifiability has been studied in continuous time settings, and recent progress has been made on parameter estimation even in weakly identifiable or nonidentifiable settings (see Section 4.5).

Question 7.1. Is there a particularly natural notion of identifiability in the discrete-time setting? If so, what are general conditions that would guarantee that a parametrized system is (or is not) identifiable? Furthermore, which properties of a system may be estimated in weakly identifiable or nonidentifiable settings, and which inference methodologies are effective in such settings?

Recall that in [111], general sufficient conditions are given under which MLE is shown to be strongly consistent in the observational noise setting. However, even under these conditions, no finite sample properties are known. In order to get finite sample error bounds, one would also like to know about the deviations of the MLE from its average. This line of reasoning leads to the following question.

Question 7.2. In the observational noise setting (4.1)–(4.2), under which conditions on the system is it true that the MLE is asymptotically normal? More generally, what are the finite sample properites of MLE in this setting?

Even in settings where statistically efficient estimators may be known to exist, there are questions and difficult issues surrounding model misspecification. How do estimators perform in the presence of model misspecification? For some results on this topic in the context of MLE for hidden Markov models, see [45]. In general, let us state the following question.

Question 7.3. Consider an estimator θ_n based on the model assumption that there is no dynamical noise. How does θ_n perform when the true system has a small amount of dynamical noise?

In the combined noise setting of Section 6, it is still the case that the issue of parameter inference has not been satisfactorily resolved. Certainly any filtering method may be trivially extended to a parameter estimation algorithm by extending the state space to include the parameters, but in such cases the degeneracy of the extended system typically causes the filtering methods to fail. Furthermore, general statistically efficient methods may be available (see, for example, [8]), but such methods are computationally intractable on large problems. Let us paraphrase a question in [47].

Question 7.4. Under what conditions on the model are there both statistically and computationally efficient algorithms for parameter estimation in the general state space setting? What theoretical guarantees can be given to justify such algorithms?

Compared to point estimation, much less is known about confidence intervals and uncertainty estimation for dynamical systems. On the frequentist side, the paper [16], shows asymptotic normality of the MLE for finite state HMMs. A few recent papers [42, 43, 46, 74] show asymptotic normality of the estimates of HMMs under increasingly general conditions. For statements regarding asymptotic normality of some estimation schemes in some ODE settings, see [138, 140, 174]. Bayesian methods, of course, automatically yield uncertainty intervals. However, not much is rigorously known about the coverage properties of Bayesian estimators for dynamical systems. On the contrary, uncertainty estimation for SDE models is well developed in both frequentist and Bayesian literature as mentioned in Section 6.5. To summarize, producing interval estimates with the right coverage is an important open area where much further work needs to be done and we state this as an open problem.

Question 7.5. Identify and develop easily verifiable conditions for which central limit theorems hold for various estimators for dynamical systems. Also derive conditions under which the posterior distribution concentrates around the true parameter at an optimal rate.

The range of applications of statistical inference methods for deterministic dynamical systems seems to be increasing rapidly. These systems present significant new challenges, since the deterministic systems may have very long-range dependency structures. It would be a significant breakthrough if one could develop asymptotically consistent algorithms for parameter estimation; moreover, one would like to have finite-size sample bounds on the accuracy of these algo-

rithms. Given the difficulty of dealing with the long-range dependencies present in general in the observational noise model, it appears likely that the traditional methods of parameter inference may not work particularly well in this setting, and therefore new ideas and methods should be developed.

One possible approach would be to consider a weakened notion of consistency. For example, one could consider a parameter estimation method to be consistent if it returns a set of plausible parameters that asymptotically contains the true parameter. Such weakened notions of consistency might be necessary for providing some theoretical justification of parameter estimation algorithms when achieving strong consistency appears out of reach.

Let us close by mentioning once again a recent development in the field of dynamical systems and ergodic theory that might be useful in obtaining asymptotic results in the absence of dynamical noise. The concentration inequalities mentioned at the end of Section 3.3 provide a powerful method for obtaining finite sample error bounds for a wide class of statistical estimators for a wide class of dynamical systems. One might hope that these concentration inequalities can be used to get rigorous error bounds for parameter estimation algorithms.

Acknowledgments

The authors would like to thank Andrew Nobel, Ramon van Handel, John Harer, Konstantin Mischaikow, Christian Robert, Mark Girolami and Andrew Stuart for discussions, comments and help with references. Additionally, the authors thank the editors and anonymous referees of the paper for offering helpful suggestions and references. SM and KM would like to acknowledge AFOSR FA9550-10-1-0436 and NSF DMS-1045153 for partial support. SM would also like to acknowledge NSF CCF-1049290 for partial support. NSP would like to thank NSF for partial support through the grant NSF DMS-1107070.

References

[1] Abarbanel, H. D. I. (1996). *Analysis of observed chaotic data. Institute for Nonlinear Science*. Springer-Verlag, New York. MR1363486

[2] Abarbanel, H. D. I., Creveling, D. R., Farsian, R. and Kostuk, M. (2009). Dynamical state and parameter estimation. *SIAM J. Appl. Dyn. Syst.* **8** 1341–1381. MR2559166

[3] Abarbanel, H. D. I., Creveling, D. R. and Jeanne, J. M. (2008). Estimation of parameters in nonlinear systems using balanced synchronization. *Phys. Rev. E (3)* **77** 016208, 14. MR2448169

[4] Adams, T. M. and Nobel, A. B. (2001). Finitary reconstruction of a measure preserving transformation. *Israel J. Math.* **126** 309–326. MR1882042

[5] Albert, R. (2007). Network Inference, Analysis, and Modeling in Systems Biology. *The Plant Cell* **19** 3327–3338.

[6] Anderson, B. D. O. and Moore, J. B. (1979). *Optimal Filtering*. Englewood Cliffs.
[7] Anderson, D. F. and Higham (2012). Multilevel Monte Carlo for continuous time Markov chains, with applications in biochemical kinetics. *SIAM: Multiscale Modeling and Simulation* **10** 146–179. MR2902602
[8] Andrieu, C., Doucet, A. and Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** 269–342. MR2758115
[9] Barreira, L. and Pesin, Y. (2006). Smooth ergodic theory and nonuniformly hyperbolic dynamics. In *Handbook of dynamical systems. Vol. 1B* 57–263. Elsevier B. V., Amsterdam With an appendix by Omri Sarig. MR2186242
[10] Berliner, L. M. (1991). Likelihood and Bayesian prediction of chaotic systems. *J. Amer. Statist. Assoc.* **86** 938–952. MR1146342
[11] Berliner, L. M. (1992). Statistics, probability and chaos. *Statist. Sci.* **7** 69–122. With discussion and a rejoinder by the author. MR1173418
[12] Beskos, A., Papaspiliopoulos, O., Roberts, G. O. and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 333–382. MR2278331
[13] Beskos, A. and Roberts, G. O. (2005). Exact simulation of diffusions. *The Annals of Applied Probability* **15** 2422–2444. MR2187299
[14] Bezruchko, B. P. and Smirnov, D. A. (2010). *Extracting knowledge from time series. Springer Series in Synergetics*. Springer, Heidelberg. An introduction to nonlinear empirical modeling. MR2767880
[15] Bhadra, A., Ionides, E. L., Laneri, K., Pascual, M., Bouma, M. and Dhiman, R. C. (2011). Malaria in Northwest India: Data analysis via partially observed stochastic differential equation models driven by Lévy noise. *Journal of the American Statistical Association* **106** 440–451. MR2866974
[16] Bickel, P. J., Ritov, Y. and Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist.* **26** 1614–1635. MR1647705
[17] Bishwal, J. P. N. (2008). *Parameter estimation in stochastic differential equations*. Springer. MR2360279
[18] Boyle, M. and Petersen, K. (2011). Hidden Markov processes in the context of symbolic dynamics. In *Entropy of hidden Markov processes and connections to dynamical systems. London Math. Soc. Lecture Note Ser.* **385** 5–71. Cambridge Univ. Press, Cambridge. MR2866664
[19] Boys, R. J., Wilkinson, D. J. and Kirkwood, T. B. L. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Stat. Comput.* **18** 125–135. MR2390814
[20] Brin, M. and Stuck, G. (2002). *Introduction to dynamical systems*. Cambridge University Press, Cambridge. MR1963683
[21] Burgers, G., Jan van Leeuwen, P. and Evensen, G. (1998). Analysis

Scheme in the Ensemble Kalman Filter. *Mon. Wea. Rev.* **126** 1719–1724.

[22] Campbell, D. and Lele, S. (2014). An ANOVA test for parameter estimability using data cloning with application to statistical inference for dynamic systems. *Computational Statistics & Data Analysis* **70** 257–267. MR3125492

[23] Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement error in nonlinear models. Monographs on Statistics and Applied Probability* **63**. Chapman & Hall, London. MR1630517

[24] Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models*, second ed. *Monographs on Statistics and Applied Probability* **105**. Chapman & Hall/CRC, Boca Raton, FL A modern perspective. MR2243417

[25] Chan, K.-S. and Tong, H. (2001). *Chaos: a statistical perspective. Springer Series in Statistics*. Springer-Verlag, New York. MR1851668

[26] Chatterjee, S. and Yilmaz, M. R. (1992). Chaos, fractals and statistics. *Statist. Sci.* **7** 49–68. MR1173417

[27] Chazottes, J. R. (2012). Fluctuations of observables in dynamical systems: from limit theorems to concentration inequalities. arXiv:1201.3833v1. MR3379397

[28] Chazottes, J. R., Collet, P., Redig, F. and Verbitskiy, E. (2009). A concentration inequality for interval maps with an indifferent fixed point. *Ergodic Theory Dynam. Systems* **29** 1097–1117. MR2529641

[29] Chazottes, J. R., Collet, P. and Schmitt, B. (2005). Devroye inequality for a class of non-uniformly hyperbolic dynamical systems. *Nonlinearity* **18** 2323–2340. MR2166315

[30] Chazottes, J. R., Collet, P. and Schmitt, B. (2005). Statistical consequences of the Devroye inequality for processes. Applications to a class of non-uniformly hyperbolic dynamical systems. *Nonlinearity* **18** 2341–2364. MR2165706

[31] Chazottes, J. R. and Gouezel, S. (2011). Optimal concentration inequalities for dynamical systems. arXiv:1111.0849v1. MR2993935

[32] Chen, K. C. C., Csikasz-Nagy, A., Gyorffy, B., Val, J., Novak, B. and Tyson, J. J. (2000). Kinetic Analysis of a Molecular Model of the Budding Yeast Cell Cycle. *Molecular Biology of the Cell* **11** 369–391.

[33] Chernov, N. (2002). Invariant measures for hyperbolic dynamical systems. In *Handbook of dynamical systems, Vol. 1A* 321–407. North-Holland, Amsterdam. MR1928521

[34] Christensen, N., Meyer, R., Knox, L. and Luey, B. (2001). Bayesian methods for cosmological parameter estimation from cosmic microwave background measurements. *Classical and Quantum Gravity* **18** 2677.

[35] Coulson, T., Rohani, P. and Pascual, M. (2004). Skeletons, noise and population growth: the end of an old debate? *Trends in Ecology & Evolution* **19** 359–364.

[36] Dattner, I. and Klaassen, C. A. (2013). Estimation in Systems of Ordinary Differential Equations Linear in the Parameters. *arXiv preprint arXiv:1305.4126*.

[37] Davies, M. (1994). Noise reduction schemes for chaotic time series. *Phys. D* **79** 174–192. MR1306461
[38] Dean, T. A., Singh, S. S., Jasra, A. and Peters, G. W. (2011). Parameter estimation for hidden Markov models with intractable likelihoods. arXiv:1103.5399v1. MR3277033
[39] Del Moral, P. (2004). *Feynman-Kac formulae. Probability and its Applications (New York)*. Springer-Verlag, New York. Genealogical and interacting particle systems with applications. MR2044973
[40] Del Moral, P., Doucet, A. and Jasra, A. (2011). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing* 1–12.
[41] Del Moral, P., Doucet, A. and Singh, S. (2010). Forward Smoothing using Sequential Monte Carlo. arXiv:1012.5390v1.
[42] Douc, R. and Matias, C. (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli* **7** 381–420. MR1836737
[43] Douc, R. and Moulines, E. (2011). Asymptotic properties of the maximum likelihood estimation in misspecified Hidden Markov models. arXiv:1110.0356v1. MR3097617
[44] Douc, R., Moulines, E., Olsson, J. and van Handel, R. (2011). Consistency of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.* **39** 474–513. MR2797854
[45] Douc, R., Moulines, E. et al. (2012). Asymptotic properties of the maximum likelihood estimation in misspecified hidden Markov models. *The Annals of Statistics* **40** 2697–2732. MR3097617
[46] Douc, R., Moulines, É. and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.* **32** 2254–2304. MR2102510
[47] Doucet, A. and Johansen, A. (2011). *A tutorial on particle filtering and smoothing: fifteen years later* 8.2. Oxford University Press. MR2884612
[48] Durbin, J. and Koopman, S. J. (2001). *Time series analysis by state space methods. Oxford Statistical Science Series* **24**. Oxford University Press, Oxford. MR1856951
[49] Eckmann, J. P. and Ruelle, D. (1985). Ergodic theory of chaos and strange attractors. *Rev. Modern Phys.* **57** 617–656. MR0800052
[50] Eubank, R. L. (2006). *A Kalman filter primer. Statistics: Textbooks and Monographs* **186**. Chapman & Hall/CRC, Boca Raton, FL. MR2193537
[51] Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99** 10143–10162.
[52] Evensen, G. (2003). The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics* **53** 343–367.
[53] Fearnhead, P. (2002). Markov Chain Monte Carlo, Sufficient Statistics, and Particle Filters. *Journal of Computational and Graphical Statistics* **11** pp. 848–862. MR1951601
[54] Fearnhead, P. and Prangle, D. (2012). Constructing summary statis-

tics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 419–474. MR2925370

[55] Fox, E. B., Sudderth, E. B., Jordan, M. I. and Willsky, A. S. (2010). Bayesian Nonparametric Methods for Learning Markov Switching Processes. *Signal Processing Magazine, IEEE* **27** 43–54.

[56] Fox, E., Sudderth, E. B., Jordan, M. I. and Willsky, A. S. (2011). Bayesian Nonparametric Inference of Switching Dynamic Linear Models. *Signal Processing, IEEE Transactions on* **59** 1569–1585.

[57] Fuller, W. A. (2006). *Measurement error models. Wiley Series in Probability and Statistics*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. Reprint of the 1987 original, Wiley-Interscience Paperback Series. MR2301581

[58] Gao, J., Sultan, H., Hu, J. and Tung, W.-W. (2010). Denoising Nonlinear Time Series by Adaptive Filtering and Wavelet Shrinkage: A Comparison. *Signal Processing Letters, IEEE* **17** 237–240.

[59] Gelman, A., Bois, F. and Jiang, J. (1996). Physiological Pharmacokinetic Analysis Using Population Modeling and Informative Prior Distributions. *Journal of the American Statistical Association* **91** 1400–1412.

[60] Girolami, M. (2008). Bayesian inference for differential equations. *Theoretical Computer Science* **408** 4–16. MR2460604

[61] Graczyk, J. and Świątek, G. (1997). Generic hyperbolicity in the logistic family. *Ann. of Math. (2)* **146** 1–52. MR1469316

[62] Gugushvili, S. and Klaassen, C. (2012). $\sqrt{n}$-consistent parameter estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing. *Bernoulli* **18** 1061–1098. MR2948913

[63] Gutman, Y. and Hochman, M. (2008). On processes which cannot be distinguished by finite observation. *Israel J. Math.* **164** 265–284. MR2391149

[64] Hamacher, K. (2012). Resilience to Leaking – Dynamic Systems Modeling of Information Security. *PLoS ONE* **7** e49804.

[65] Hamilton, J. D. (1994). *Time-series analysis*, 1 ed. Princeton Univerity Press. MR1278033

[66] Himmelblau, D., Jones, C. and Bischoff, K. (1967). Determination of rate constants for complex kinetics models. *Industrial & Engineering Chemistry Fundamentals* **6** 539–543.

[67] Iacus, S. M. (2008). *Simulation and inference for stochastic differential equations: with R examples*. Springer. MR2410254

[68] Ionides, E. L., Bhadra, A., Atchadé, Y. and King, A. (2011). Iterated filtering. *Ann. Statist.* **39** 1776–1802. MR2850220

[69] Isham, V. (1993). Statistical aspects of chaos: a review. In *Networks and chaos—statistical and probabilistic aspects. Monogr. Statist. Appl. Probab.* **50** 124–200. MR1314654

[70] Jaeger, L. and Kantz, H. (1996). Unbiased reconstruction of the dynamics underlying a noisy chaotic time series. *Chaos* **6** 440–450.

[71] Jakobson, M. V. (1981). Absolutely continuous invariant measures for

one-parameter families of one-dimensional maps. *Comm. Math. Phys.* **81** 39–88. MR0630331

[72] Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*. Academic Press.

[73] Jensen, J. L. (1993). Chaotic dynamical systems with a view towards statistics: a review. In *Networks and chaos—statistical and probabilistic aspects. Monogr. Statist. Appl. Probab.* **50** 201–250. MR1314655

[74] Jensen, J. L. and Petersen, N. V. (1999). Asymptotic Normality of the Maximum Likelihood Estimator in State Space Models. *The Annals of Statistics* **27** pp. 514-535. MR1714719

[75] Judd, K. (2003). Chaotic-time-series reconstruction by the Bayesian paradigm: Right results by wrong methods. *Phys. Rev. E* **67** 026212.

[76] Judd, K. (2003). Nonlinear state estimation, indistinguishable states, and the extended Kalman filter. *Phys. D* **183** 273–281. MR2006638

[77] Judd, K. (2007). Failure of maximum likelihood methods for chaotic dynamical systems. *Phys. Rev. E* **75** 036210.

[78] Judd, K. (2008). Shadowing Pseudo-Orbits and Gradient Descent Noise Reduction. *Journal of Nonlinear Science* **18** 57–74. MR2387132

[79] Judd, K. and Smith, L. (2001). Indistinguishable states. I. Perfect model scenario. *Phys. D* **151** 125–141. MR1834043

[80] Judd, K. and Smith, L. A. (2004). Indistinguishable states II: The imperfect model scenario. *Physica D: Nonlinear Phenomena* **196** 224–242. MR2090352

[81] Julier, S. J. and Uhlmann, J. K. (1996). A general method for approximating nonlinear transformations of probability distributions Technical Report, Department of Engineering Science, Oxford University.

[82] Julier, S. J. and Uhlmann, J. K. (1997). A new extension of the Kalman filter to nonlinear systems. In *Proc. of AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulation, and Controls*.

[83] Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering* **82** 35–45.

[84] Kantas, N., Doucet, A., Singh, S. S. and Maciejowski, J. M. (2009). An overview of sequential Monte Carlo methods for parameter estimation in general state-space models. In *Proceedings of the IFAC System Identification Meeting*.

[85] Kantz, H. and Jaeger, L. (1997). Improved cost functions for modelling of noisy chaotic time series. *Physica D: Nonlinear Phenomena* **109** 59–69.

[86] Kantz, H. and Schreiber, T. (2004). *Nonlinear time series analysis*, Second ed. Cambridge University Press, Cambridge. MR2040330

[87] Katok, A. and Hasselblatt, B. (1995). *Introduction to the modern theory of dynamical systems. Encyclopedia of Mathematics and its Applications* **54**. Cambridge University Press, Cambridge. With a supplementary chapter by Katok and Leonardo Mendoza. MR1326374

[88] Kenett, R. S., Harel, A. and Ruggeri, F. (2009). Controlling the

Usability of Web Services. *International Journal of Software Engineering and Knowledge Engineering* 627–651.

[89] Kifer, Y. (1988). *Random perturbations of dynamical systems. Progress in Probability and Statistics* **16**. Birkhäuser Boston Inc., Boston, MA. MR1015933

[90] Kifer, Y. and Liu, P.-D. (2006). Random dynamics. In *Handbook of dynamical systems. Vol. 1B* 379–499. Elsevier B. V., Amsterdam. MR2186245

[91] King, A. A., Costantino, R., Cushing, J., Henson, S. M., Desharnais, R. A. and Dennis, B. (2004). Anatomy of a chaotic attractor: subtle model-predicted patterns revealed in population data. *Proceedings of the National Academy of Sciences* **101** 408–413.

[92] Kostelich, E. and Schreiber, T. (1993). Noise reduction schemes for chaotic time-series data: a survey of common methods. *Phys. Rev. E* **48** 1752–1763. MR1377916

[93] Kostelich, E. J. (1992). Problems in estimating dynamics from data. *Physica D: Nonlinear Phenomena* **58** 138–152. MR1188246

[94] Kostelich, E. J. and Yorke, J. A. (1990). Noise reduction: finding the simplest dynamical system consistent with the data. *Phys. D* **41** 183–196. MR1049125

[95] Kou, S., Olding, B. P., Lysy, M. and Liu, J. S. (2012). A multiresolution method for parameter estimation of diffusion processes. *Journal of the American Statistical Association* **107** 1558–1574. MR3036416

[96] Kou, S., Sunney Xie, X. and Liu, J. S. (2005). Bayesian analysis of single-molecule experimental data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54** 469–506. MR2137252

[97] Kreutz, C., Raue, A., Kaschek, D. and Timmer, J. (2013). Profile likelihood in systems biology. *FEBS Journal* **280** 2564–2571.

[98] Kundur, D., Feng, X., Mashayekh, S., Liu, S., Zourntos, T. and Butler-Purry, K. L. (2011). Towards modelling the impact of cyber attacks on a smart grid. *International Journal of Security and Networks* **6** 2–13.

[99] Lalley, S. P. (1999). Beneath the noise, chaos. *Ann. Statist.* **27** 461–479. MR1714721

[100] Lalley, S. P. (2001). Removing the noise from chaos plus noise. In *Nonlinear dynamics and statistics (Cambridge, 1998)* 233–244. Birkhäuser Boston, Boston, MA. MR1937487

[101] Lalley, S. P. and Nobel, A. B. (2006). Denoising deterministic time series. *Dyn. Partial Differ. Equ.* **3** 259–279. MR2271730

[102] Lopes, H. F. and Tsay, R. S. (2011). Particle filters and Bayesian inference in financial econometrics. *Journal of Forecasting* **30** 168–209. MR2758809

[103] Lyubich, M. (1994). Combinatorics, geometry and attractors of quasi-quadratic maps. *Ann. of Math. (2)* **140** 347–404. MR1298717

[104] Lyubich, M. (2002). Almost every real quadratic map is either regular or stochastic. *Ann. of Math. (2)* **156** 1–78. MR1935840

[105] Manjunath, G., Sivaji Ganesh, S. and Anand, G. V. (2009). Topology-based denoising of chaos. *Dyn. Syst.* **24** 501–516. MR2573001

[106] Manjunath, G., Sivaji Ganesh, S. and Anand, G. V. (2010). Denoising signals corrupted by chaotic noise. *Commun. Nonlinear Sci. Numer. Simul.* **15** 3988–3997. MR2652670

[107] Manolopoulou, I., Matheu, M. P., Cahalan, M. D., West, M. and Kepler, T. B. (2012). Bayesian Spatio-Dynamic Modeling in Cell Motility Studies: Learning Nonlinear Taxic Fields Guiding the Immune Response. *Journal of the American Statistical Association* **107** 855–865. MR3010872

[108] Marin, J. M., Pillai, N. S., Robert, C. P. and Rousseau, J. (2011). Relevant statistics for Bayesian model choice. *ArXiv e-prints*. MR3271169

[109] Marin J. M., R. C. P. Pudlo P. and Ryder, R. Approximate Bayesian Computational methods. *Statistics and Computing* **2** 289–291.

[110] Maybhate, A. and Amritkar, R. E. (1999). Use of synchronization and adaptive control in parameter estimation from a time series. *Phys. Rev. E* **59** 284–293.

[111] McGoff, K., Mukherjee, S., Nobel, A. and Pillai, N. Consistency of maximum likelihood estimation for some dynamical systems. *Annals of Statistics*. to appear. MR3285598

[112] McKinley, T., Cook, A. R. and Deardon, R. (2009). Inference in Epidemic Models without Likelihoods. *The International Journal of Biostatistics* **5** 24. MR2533810

[113] McLean, K. A., Wu, S. and McAuley, K. B. (2012). Mean-squared-error methods for selecting optimal parameter subsets for estimation. *Industrial & Engineering Chemistry Research* **51** 6105–6115.

[114] McSharry, P. E. and Smith, L. A. (1999). Better Nonlinear Models from Noisy Data: Attractors with Maximum Likelihood. *Phys. Rev. Lett.* **83** 4285–4288.

[115] Mees, A. I., ed. (2001). *Nonlinear dynamics and statistics*. Birkhäuser Boston Inc., Boston, MA. Selected papers from the workshop held at Cambridge University, Cambridge, September 1998. MR1936437

[116] Meyer, R. and Christensen, N. (2000). Bayesian reconstruction of chaotic dynamical systems. *Physical Review E.* **62**.

[117] Meyn, S. and Tweedie, R. L. (2009). *Markov chains and stochastic stability*, Second ed. Cambridge University Press, Cambridge. With a prologue by Peter W. Glynn. MR2509253

[118] Miao, H., Xia, X., Perelson, A. and Wu, H. (2011). On Identifiability of Nonlinear ODE Models and Applications in Viral Dynamics. *SIAM Review* **53** 3–39. MR2785878

[119] Mukherjee, C. (2011). Bayesian Modelling and Computation in Dynamic and Spatial Systems PhD thesis, Duke University, Durham, North Carolina. MR2926830

[120] Mukherjee, C. and West, M. (2009). Sequential Monte Carlo in model comparison: Example in cellular dynamics in systems biology. In *JSM Proceedings,Section on Bayesian Statistical Science. Alexandria, VA: Ameri-*

can Statistical Association 1274–1287.

[121] Nakamura, T., Hirata, Y., Judd, K., Kilminster, D. and Small, M. (2007). Improved parameter estimation from noisy time series for nonlinear dynamical systems. *Internat. J. Bifur. Chaos Appl. Sci. Engrg.* **17** 1741–1752. MR2339949

[122] Nobel, A. (2001). Consistent estimation of a dynamical map. In *Nonlinear dynamics and statistics (Cambridge, 1998)* 267–280. Birkhäuser Boston, Boston, MA. MR1937489

[123] Nobel, A. B. and Adams, T. M. (2001). Estimating a function from ergodic samples with additive noise. *IEEE Trans. Inform. Theory* **47** 2895–2902. MR1872848

[124] Olsson, J., Cappé, O., Douc, R. and Moulines, E. (2008). Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli* **14** 155–179. MR2401658

[125] Ornstein, D. and Weiss, B. (2007). Entropy is the only finitely observable invariant. *J. Mod. Dyn.* **1** 93–105. MR2261073

[126] Ornstein, D. S. and Weiss, B. (1990). How sampling reveals a process. *Ann. Probab.* **18** 905–930. MR1062052

[127] Ornstein, D. S. and Weiss, B. (1991). Statistical properties of chaotic systems. *Bull. Amer. Math. Soc. (N.S.)* **24** 11–116. With an appendix by David Fried. MR1023980

[128] Packard, N. H., Crutchfield, J. P., Farmer, J. D. and Shaw, R. S. (1980). Geometry from a time series. *Phys. Rev. Lett.* **45** 712–715.

[129] Parlitz, U. (1996). Estimating Model Parameters from Time Series by Autosynchronization. *Phys. Rev. Lett.* **76** 1232–1235.

[130] Peters, G. W., Wüthrich, M. V. and Shevchenko, P. V. (2010). Chain ladder method: Bayesian bootstrap versus classical bootstrap. *Insurance: Mathematics and Economics* **47** 36–51. MR2675675

[131] Petersen, K. (1989). *Ergodic theory. Cambridge Studies in Advanced Mathematics* **2**. Cambridge University Press, Cambridge. Corrected reprint of the 1983 original. MR1073173

[132] Pievatolo, A., Ruggeri, F. and Soyer, R. (2012). A Bayesian hidden Markov model for imperfect debugging. *Rel. Eng. & Sys. Safety* 11–21.

[133] Pisarenko, V. F. and Sornette, D. (2004). Statistical methods of parameter estimation for deterministically chaotic time series. *Phys. Rev. E* **69** 036122. MR2096393

[134] Pole, A., West, M. and Harrison, P. J. (1994). *Applied Bayesian Forecasting & Time Series Analysis*. Chapman-Hall.

[135] Pollard, D. and Radchenko, P. (2006). Nonlinear least-squares estimation. *J. Multivariate Anal.* **97** 548–562. MR2234037

[136] Poyiadjis, G., Doucet, A. and Singh, S. S. (2011). Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. **98** 65–80. MR2804210

[137] Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*

16 1791–1798.

[138] Qi, X. and Zhao, H. (2010). Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *Ann. Statist.* **38** 435–481. MR2589327

[139] Quinn, J. C., Bryant, P. H., Creveling, D. R., Klein, S. R. and Abarbanel, H. D. I. (2009). Parameter and state estimation of experimental chaotic systems using synchronization. *Phys. Rev. E (3)* **80** 016201, 17. MR2552045

[140] Ramsay, J. O., Hooker, G., Campbell, D. and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 741–796. With discussions and a reply by the authors. MR2368570

[141] Ratmann, O., Andrieu, C., Wiuf, C. and Richardson, S. (2009). Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences* **106** 10576–10581.

[142] Raue, A., Becker, V., Klingmüller, U. and Timmer, J. (2010). Identifiability and observability analysis for experimental design in nonlinear dynamical models. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **20** 045105.

[143] Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U. and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25** 1923–1929.

[144] Ridout, D. and Judd, K. (2002). Convergence properties of gradient descent noise reduction. *Physica D: Nonlinear Phenomena* **165** 26–47. MR1910616

[145] Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*, second ed. *Springer Texts in Statistics*. Springer-Verlag, New York. MR2080278

[146] Robert, C. P., Cornuet, J. M., Marin, J. M. and Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computational (ABC) model choice. *PNAS* **108** 15112–15117.

[147] Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12** 1151–1172. MR0760681

[148] Sauer, T. (1992). A noise reduction method for signals from nonlinear systems. *Phys. D* **58** 193–201. Interpretation of time series from nonlinear systems (Warwick, 1991). MR1188249

[149] Sauer, T., Yorke, J. A. and Casdagli, M. (1991). Embedology. *J. Statist. Phys.* **65** 579–616. MR1137425

[150] Shalizi, C. R. (2009). Dynamics of Bayesian updating with dependent data and misspecified models. *Electron. J. Stat.* **3** 1039–1074. MR2557128

[151] Simmons Kovacs, L., Mayhew, M. B., Orlando, D. A., Jin, Y., Li, Q., Huang, C., Reed, S. I., Mukherjee, S. and Haase, S. B. (2012). Cyclin-Dependent Kinases Are Regulators and Effectors of Os-

cillations Driven by a Transcription Factor Network. *Molecular Cell* **45** 669–679.

[152] Sisson, S. A., Fan, Y. and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **104** 1760–1765. MR2301870

[153] Smirnov, D. A., Vlaskin, V. S. and Ponomarenko, V. I. (2005). Estimation of parameters in one-dimensional maps from noisy chaotic time series. *Phys. Lett. A* **336** 448–458. MR2118773

[154] Srinath, S. and Gunawan, R. (2010). Parameter identifiability of power-law biochemical system models. *Journal of biotechnology* **149** 132–140.

[155] Stark, J., Broomhead, D. S., Davies, M. E. and Huke, J. (2003). Delay embeddings for forced systems. II. Stochastic forcing. *J. Nonlinear Sci.* **13** 519–577. MR2020239

[156] Stemler, T. and Judd, K. (2009). A guide to using shadowing filters for forecasting and state estimation. *Phys. D* **238** 1260–1273. MR2532407

[157] Storvik, G. (2002). Particle filters for state-space models with the presence of unknown static parameters. *Signal Processing, IEEE Transactions on* **50** 281–289.

[158] T., G. D. (2012). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem* **81** 2340–2361.

[159] Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980 (Coventry, 1979/1980). Lecture Notes in Math.* **898** 366–381. Springer, Berlin. MR0654900

[160] Tavare, S., Balding, D. J., Griffiths, R. C. and Donnelly, P. (1997). Inferring Coalescence Times from DNA Sequence Data. *Genetics* **145** 505–518.

[161] Tong, H. (1990). *Nonlinear time series. Oxford Statistical Science Series* **6**. The Clarendon Press Oxford University Press, New York. A dynamical system approach, With an appendix by K. S. Chan, Oxford Science Publications. MR1079320

[162] Toni, T. and Strumpf, M. P. H. (2010). Simulation-based model selection for dynamical systems in population and systems biology. *Bioinformatics* **26** 104–110.

[163] Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6** 187–202.

[164] Turchin, P. and Ellner, S. P. (2000). Modelling time-series data. In *Chaos in real data* 33–48. Springer.

[165] Vanlier, J., Tiemann, C. A., Hilbers, P. A. and van Riel, N. A. (2012). An integrated strategy for prediction uncertainty analysis. *Bioinformatics* **28** 1130–1135.

[166] Varah, J. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing* **3** 28–46. MR0651865

[167] Voss, H. U., Timmer, J. and Kurths, J. (2004). Nonlinear dynamical system identification from uncertain and indirect measurements. *Internat. J. Bifur. Chaos Appl. Sci. Engrg.* **14** 1905–1933. MR2076173

[168] Walters, P. (1982). *An introduction to ergodic theory. Graduate Texts in Mathematics* **79**. Springer-Verlag, New York. MR0648108

[169] West, M. and Harrison, P. J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed. Springer Verlag. MR1482232

[170] Wilkinson, D. J. (2009). Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat. Rev. Genet.* **10** 122–133.

[171] Wilkinson, D. J. (2011). *Stochastic modelling for systems biology* **44**. CRC press. MR2222876

[172] Wilkinson, R. D. (2008). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. arXiv:0811.3355v1. MR3071024

[173] Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466** 1102–1104.

[174] Xue, H., Miao, H. and Wu, H. (2010). Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. *Ann. Statist.* **38** 2351–2387. MR2676892

[175] Xun, X., Cao, J., Mallick, B., Maity, A. and Carroll, R. J. (2013). Parameter estimation of partial differential equation models. *Journal of the American Statistical Association* **108** 1009–1020. MR3174680

[176] Young, L.-S. (1998). Statistical properties of dynamical systems with some hyperbolicity. *Ann. of Math. (2)* **147** 585–650. MR1637655

[177] Young, L.-S. (2002). What are SRB measures, and which dynamical systems have them? *J. Statist. Phys.* **108** 733–754. Dedicated to David Ruelle and Yasha Sinai on the occasion of their 65th birthdays. MR1933431

[178] Yu, W., Chen, G., Cao, J., Lü, J. and Parlitz, U. (2007). Parameter identification of dynamical systems from time series. *Phys. Rev. E* **75** 067201.