

Parameter estimation and model selection for stochastic differential equations for biological growth

F. Baltazar-Larios^a, F.J. Delgado-Vences^b and A. Ornelas Vargas^c

^aFacultad de Ciencias, Universidad Nacional Autónoma de México, México; ^b Conacyt Research Fellow, Instituto de Matemáticas, Universidad Nacional Autónoma de México, Oaxaca, México; Conacyt Research Fellow, Centro Interdisciplinario de Ciencias Marinas, Instituto Politécnico Nacional, La Paz, México.

ARTICLE HISTORY

Compiled January 23, 2023

ABSTRACT

In this paper, we consider stochastic versions of three classical growth models given by ordinary differential equations (ODEs). Indeed we use stochastic versions of Von Bertalanffy, Gompertz, and Logistic differential equations as models. We assume that each stochastic differential equation (SDE) has some crucial parameters in the drift to be estimated and we use the Maximum Likelihood Estimator (MLE) to estimate them. For estimating the diffusion parameter, we use the MLE for two cases and the quadratic variation of the data for one of the SDEs. We apply the Akaike information criterion (AIC) to choose the best model for the simulated data. We consider that the AIC is a function of the drift parameter. We present a simulation study to validate our selection method.

The proposed methodology could be applied to datasets with continuous and discrete observations, but also with highly sparse data. Indeed, we can use this method even in the extreme case where we have observed only one point for each path, under the condition that we observed a sufficient number of trajectories. For the last two cases, the data can be viewed as incomplete observations of a model with a tractable likelihood function; then, we propose a version of the Expectation Maximization (EM) algorithm to estimate these parameters. This type of datasets typically appears in fishery, for instance.

Keywords: Stochastic differential equations; Maximum Likelihood estimation; EM algorithm; model selection; AIC

1. Introduction

The main motivation for this work comes from problems where it is necessary to fit SDEs to modeling growth biological data; for instance in marine biology, ecology (see [8]), oncology (see [18]), or in paleontology (to model sclerochronological parameters of shell growth [16]). Given that real systems cannot be completely isolated from their environments and, therefore, always experience external stochastic forces, then the use of SDEs as models is justified and preferred to (ODEs). SDEs have been used in several fields such as Finance, econometrics, population systems, ecology, etc.; thus, fitting SDEs to actual data has wide interest.

Fitting differential equations is a classic example in inverse problems, and has been treated by many authors with several approaches (see for instance [11], [13], or [15]) and, in particular, fitting ODEs have been deeply studied. One typical situation in these problems is when fitting ODEs is equivalent to adjusting only a fixed number of unknown and constant parameters. Motivated for this idea, we will assume that we have this case, meaning we assume that every SDE has some crucial parameters that we focus on estimating (see [7] or [22]).

In this work, we consider the problem of estimating the parameters using the available data to fit each SDEs and to propose a criterion to choose the best fit. There exists theory to do this estimation and, depending on the observations data available, could be applied to solve the problem. However, depending on the available data it is possible to use one of the existing methods to estimate parameters (see, for instance, [12], [19]). The three scenarios we consider in this work are: a continuous observation of the solutions, discretized observations, and only one observed measurement for a nice number of individuals. The last scenario has the advantage that the data is a collection of several individuals from the same population, which allows us to mix the observation to simulate discretized observations. This type of data appear usually in the fishery. Thus, in this work we fit SDEs to the three scenarios of observations described above.

A statistical model is a useful tool for the description and forecast of a stochastic (or deterministic) system. We could roughly say that a statistical model is a simplification of a complex reality. We could think that the complex reality is a “true” model and one of the goals of mathematicians and statisticians is to approximate (in some sense) this “true” model. Depending on the problem, we might assume that the “true” model is contained within a set of models under our consideration [1]. To choose the best model assuming a given dataset could be done using several, the most used is Akaike’s information criterion, known as AIC, for evaluating the constructed models. AIC has been applied successfully to linear regression, time series, etc.

The stochastic models we focus on in this paper are Gompertz, Von Bertalanffy, and Logistic stochastic differential equations. We consider that some parameters in each SDE are constants but unknown, usually, the drift parameter will represent the intrinsic growth rate of the model; thus, we are interested in estimating this parameter. Given the data, we are interested in estimating the parameters for each SDE and we apply the AIC to each of the three SDEs which provides a criterion to choose the best model that fits to data.

Parameter estimation for SDEs using the MLE method (continuous observation), the EM algorithm, the Ozaki method, and Bayesian methods (discrete observation), has been studied in [19] and [12]. In this work, we calculate the likelihood function of the transition probabilities for the parameter of Gompertz and Von Bertalanffy models and obtain the MLEs for the parameters in the continuous case and use the EM algorithm to obtain the corresponding MLEs for the case of incomplete information (discrete observation). For the Logistic model, we follow the method for estimating the parameters described in [9], where the estimation combines quadratic variation of the observations and MLE via EM algorithm. For the case where only a data set corresponding to one record for each individual, but with observations of several different individuals at different points in time, we propose a novel algorithm to simulate paths of the SDEs and obtained the corresponding parameters.

Since we know that by maximizing the log-likelihood, it is possible to obtain the estimators and we conclude that we have good estimators for Gompertz and Von Bertalanffy models. For the Logistic model, the properties of the estimators were proved in [9].

It is known that the AIC is a function of the log-likelihood, which depends on the parameters of the model. In this paper, we use the MLEs (Gompertz and Von Bertalanffy) to calculate the AIC. For the Logistic model, we assume that the log-likelihood is a function only of the drift parameters, meaning that the diffusion parameter is previously fixed, afterward we calculate the corresponding AIC. We validate this method using simulated data for the three SDEs and, then we fit the best stochastic model.

This paper is organized as follows. In Section 2 we present the SDEs considered in this manuscript, Section 3 contains the parameter estimation under three types of datasets. Section 4 presents the simulations study for the different scenarios and the numerical result of this simulation. In Section 5 we apply the AIC criteria to the SDEs and we present the results of numerical simulations. Section 6 includes the conclusions and final remarks of the work.

2. Three stochastic growth models

We consider the one-dimensional, time-homogeneous SDE

$$dX_t = \alpha_\theta(X_t)dt + \beta_\theta(X_t)dW_t, \quad (1)$$

where W_t is an standard Wiener process, θ is the multidimensional parameter to be estimated and α and β are known functions such that the solution of (1) exists. We will consider three particular functions for α and β and, therefore, we will obtain three stochastic growth models, which are the subject of this manuscript.

2.1. A stochastic Gompertz differential equation

Consider the stochastic Gompertz differential equation given by

$$\begin{aligned} dX_t &= -bX_t \log(X_t)dt + \sigma X_t dW_t, & t > 0 \\ X_0 &= x_0, \end{aligned} \quad (2)$$

i.e., we have taken $\alpha_\theta(X_t) = -bX_t \log(X_t)$, $\beta_\theta(X_t) = \sigma X_t$, and $\theta = (b, \sigma)$, where b and $\sigma > 0$ are constants. In this equation, we are setting the carrying capacity, or in our case the maximum size that can be reached by the spice, equal to 1.

The solution to (2) is obtained by applying the Itô formula to $y = \log(x)$. Indeed, we have that the stochastic process Y_t solves

$$\begin{aligned} dY_t &= d\log(X_t) = \left[-b\log(X_t) - \frac{\sigma^2}{2} \right] dt + \sigma dW_t \\ &= \left[-\frac{\sigma^2}{2} - bY_t \right] dt + \sigma dW_t, \end{aligned}$$

which we identify as the Ornstein-Uhlenbeck process for $Y_t = \log(X_t)$. This implies (see [20] for further reading) that the solution is

$$\log(X_t) = Y_t = Y_0 e^{-bt} + \frac{-\sigma^2}{2b}(1 - e^{-bt}) + \sigma \int_0^t e^{-b(t-s)} dW_s,$$

where $Y_0 = \log(X_0)$.

Thus, the solution to (2) is

$$X_t = \exp \left[\log(X_0) e^{-bt} - \frac{\sigma^2}{2b} (1 - e^{-bt}) + \sigma \int_0^t e^{-b(t-s)} dW_s \right]. \quad (3)$$

Recall that the stochastic integral is Gaussian and that if $Z \sim \mathcal{N}(m, \sigma^2)$ then $\mathbb{E}[\exp(tZ)] = e^{mt} e^{\sigma^2 t^2/2}$. Using these facts, it is not difficult to see that

$$\mathbb{E}(X_t) = \exp \left[\log(X_0) e^{-bt} - \frac{\sigma^2}{2b} (1 - e^{-bt}) + \frac{\sigma^2}{4b} (1 - e^{-2bt}) \right], \quad (4)$$

and then,

$$\lim_{t \rightarrow \infty} \mathbb{E}(X_t) = \exp \left[-\frac{\sigma^2}{4b} \right].$$

Remark 2.1. We observe that taking $\sigma = 0$ in (2) we recover a deterministic ordinary differential equation. Furthermore, using expression (4) we can, heuristically, find a solution for the deterministic version of (2) given by $\exp[C e^{-bt}]$, which agrees with the classical one.

2.2. A stochastic Von Bertalanffy differential equation

We take $\alpha_\theta(L_t) = \kappa(L_\infty - L_t)$ and $\beta_\theta(X_t) = \sigma(L_\infty - L_t)$, thus we obtained the stochastic Von Bertalanffy differential equation, which is the SDE given by

$$\begin{aligned} dL_t &= \kappa(L_\infty - L_t)dt + \sigma(L_\infty - L_t)dW_t, \quad t > 0, \\ L_0 &= l_0, \end{aligned} \quad (5)$$

where we assume L_∞ to be known, $\theta = (\kappa, \sigma)$. Here κ is interpreted as the growth coefficient, which we assume is constant but unknown, and σ is the diffusion parameter.

By applying Itô formula to the function $G_t = (L_\infty - L_t)$ we get

$$\begin{aligned} dG_t &= d(L_\infty - L_t) = -\kappa(L_\infty - L_t)dt - \sigma(L_\infty - L_t)dW_t \\ &= -\kappa G_t - \sigma G_t dW_t. \end{aligned}$$

From the last equality, we identify G as a Geometric Brownian motion with solution (see [20]) given by

$$G_t = G_0 \exp \left(\left(-\kappa - \frac{\sigma^2}{2} \right) t - \sigma W_t \right), \quad (6)$$

with $G_0 = (L_\infty - l_0)$.

From the explicit expression for G_t we get that

$$L_t = L_\infty - (L_\infty - l_0) \exp \left(\left(-\kappa - \frac{\sigma^2}{2} \right) t - \sigma W_t \right). \quad (7)$$

Note that

$$\lim_{t \rightarrow \infty} \mathbb{E}(L_t) = L_\infty.$$

2.3. A stochastic logistic differential equation

In this section, we present the SLDE which is obtained when we set $\alpha_\theta(P_t) = rP_t(1 - P_t)$ and $\sigma_\theta(P_t) = \sigma P_t$, then we have

$$dP_t = rP_t(1 - P_t)dt + \sigma P_t dW_t, \quad (8)$$

where $\sigma > 0$ and $p_0 = P_0$ is a bounded absolutely continuous random variable $p_0(\omega) : \Omega \rightarrow [a_1, a_2] \subset (0, 1)$. In the model (8) the size is dependent on the corresponding population. This equation has been studied in [19].

It is well-known that the deterministic version of the equation (8) has been a good model for several phenomena (see for instance [4] or [3] and the references therein). In our case, P_t will denote the proportional size of the individual of a given population, with p_0 being the random initial size. $r > 0$ is the intrinsic growth rate of the species, and we assume it is our interest parameter to be estimated.

The strong solution (see [14, see Th. 2.2. therein]) to (8) given by

$$P_t = \frac{f_t}{\frac{1}{p_0} + r \int_0^t f_s ds}, \quad (9)$$

where

$$f_t := \exp \left(t \left(r - \frac{\sigma^2}{2} \right) + \sigma W_t \right).$$

We have that

$$\lim_{t \rightarrow \infty} \mathbb{E} f_t = e^{rt},$$

and, we deduce that

$$\lim_{t \rightarrow \infty} P_t = 1.$$

3. Parametric estimation

In this section, we will assume that the parameter in each SDE is unknown and that the initial condition, for each SDE, is a random variable with some given density. In

addition, we assume this density is the same for all possible values of the parameter. Fixed a positive time $T > 0$ and a time interval $[0, T]$. We present algorithms to estimate the parameters of equations (2), (5), and (8) in three different scenarios. First, we assume that we observe the data at continuous times in $[0, T]$, which is unreal. The second case is when the observations are given at discrete time. Finally, when we observed only one single point for each path for a suitable number of trajectories. The last case mean that every observation represent a different member of the population we modeled. For the two last scenarios we use the EM algorithm to estimate the parameters (see [6], [10], and [17]) by completing the information; that is, we see the gap between two consecutive points as a missed information, and using diffusion bridges we want to complete the information. Here, we are assuming that the models have a tractable likelihood function. In the EM algorithm it is necessary to calculate the conditional expectation in the E-step, which is done using the approximation of diffusion bridges given in [2]. Thus, we calculate the MLE for the drift parameter in each SDE and we use the asymptotic properties of the MLE. Notice that the noise in each SDE is a multiplicative type; indeed, it is an affine function of the solution of the SDE multiplied by the Wiener process. Therefore, with the previous assumptions, the diffusion parameter can be estimated from the quadratic variation when the MLE corresponding cannot be determined.

3.1. Continuous case

For the case of continuous observation in the interval time $[0, T]$, we assume that the observation interval is divided into n sub-intervals $[t_{i-1}, t_i]$ of length $\Delta_n = \frac{T}{n}$ with $0 = t_0 < t_1 < \dots < t_n = T$. Thus, to denote that we observe the paths continuously, we assume that n is large enough such that Δ_n goes to zero.

3.1.1. MLE for a stochastic Gompertz model

To compute the MLE of parameters $\theta = (b, \sigma)$ of the SDE (2), under continuous observation in $[0, T]$, we use the solution given in (3) and we take advantage that the likelihood function has a closed form given by

$$L(y_0, y_1, \dots, y_n; b, \sigma) = \prod_{i=1}^n \left[\frac{\exp \left(-\frac{(y_i - (y_{i-1} e^{-b\Delta_n} - (\sigma^2/2b)(1 - e^{-b\Delta_n})))^2}{\sigma^2(1 - e^{-2b\Delta_n})/b} \right)}{\sqrt{2\pi\sigma^2(1 - e^{-2b\Delta_n})/b}} \right], \quad (10)$$

where $Y_{t_0} = y_0, Y_{t_1} = y_1, \dots, Y_{t_n} = y_n$. The MLEs are obtained by finding the maximum of the function (10). Thus, the MLEs are

$$\hat{b} = -\frac{c_1}{\Delta_n}, \quad (11)$$

and

$$\hat{\sigma} = \sqrt{\frac{2c_3 h \hat{b}}{1 - e^{-2\hat{b}}}}, \quad (12)$$

where

$$\begin{aligned}
c_1 &= \frac{n \sum_{i=1}^n y_i y_{i-1} - \sum_{i=1}^n y_i \sum_{i=1}^n y_{i-1}}{n \sum_{i=1}^n y_{i-1}^2 - (\sum_{i=1}^n y_{i-1})^2}, \\
c_2 &= \frac{\sum_{i=1}^n y_{i-1} - c_1 \sum_{i=1}^n y_i}{n}, \\
c_3 &= \frac{\sum_{i=1}^n (y_i - c_1 y_{i-1} + c_2)}{n}.
\end{aligned} \tag{13}$$

3.1.2. MLE for a stochastic Von Bertalanffy model

For this model, we have the observations $\{L_{t_0}, L_{t_1}, \dots, L_{t_n}\}$. Assume we know the value of L_∞ , then we can obtain the corresponding observations $G_{t_i} = L_\infty - L_{t_i}$. By (6), the conditional density function of the $\{G_t\}_{t=0}^T$ is log-normal and then the transition probabilities are

$$p_{\Delta_n}(g_i | g_{i-1}) = \frac{\exp\left(-\frac{[\log(g_i) - (\log(g_{i-1}) + (\kappa - \sigma^2/2)\Delta_n)]^2}{2\sigma^2\Delta_n}\right)}{g_i \sigma \sqrt{2\pi\Delta_n}},$$

where $G_{t_i} = g_i$, $i = 0, 1, \dots, n$. We have that the likelihood function is

$$L(g_0, g_1, \dots, g_n; \kappa, \sigma) = \prod_{i=1}^n \left[\frac{\exp\left(-\frac{[\log(g_i) - (\log(g_{i-1}) + (\kappa - \sigma^2/2)\Delta_n)]^2}{2\sigma^2\Delta_n}\right)}{g_i \sigma \sqrt{2\pi\Delta_n}} \right]. \tag{14}$$

From (14) is easy see that the maximum likelihood estimators are

$$\hat{\sigma}^2 = \frac{-a^2 + 2ab - b^2 + cn - 2dn + en}{nT}. \tag{15}$$

$$\hat{\kappa} = \frac{b-a}{T} - \frac{\hat{\sigma}^2}{2}. \tag{16}$$

where

$$\begin{aligned}
a &= \sum_{i=1}^n \log(g_i), \\
b &= \sum_{i=1}^n \log(g_{i-1}), \\
c &= \sum_{i=1}^n \log^2(g_i), \\
d &= \sum_{i=1}^n \log(g_{i-1}) \log(g_i), \\
e &= \sum_{i=1}^n \log^2(g_{i-1}).
\end{aligned} \tag{17}$$

3.1.3. A stochastic logistic model

For this model, we use the quadratic variations to obtain an estimator of σ . The MLE for r is presented in [9], where the authors proved that the MLE is strongly consistent and asymp-

totically normal. Let $P_{t_0} = p_0, P_{t_1} = p_1, \dots, P_{t_n} = p_n$ be the observations in the interval time $[0, T]$. Following [9], we have that an estimator of σ is given by

$$\hat{\sigma}^2 = \frac{2 \sum_{i=1}^n (p_i - p_{i-1})^2}{\sum_{i=1}^n (p_i^2 - p_{i-1}^2)}. \quad (18)$$

If we estimate σ using the expression (18) we have that the full log-likelihood function of r is

$$l(r) = r \frac{1}{\sigma^2} \int_0^T \frac{(1 - P_t)}{P_t} dP_t - \frac{1}{2} r^2 \frac{1}{\sigma^2} \int_0^T (1 - P_t^2) dt.$$

The MLE of r using the data is

$$\hat{r} = \frac{\int_0^T \frac{(1 - P_t)}{P_t} dP_t}{\int_0^T (1 - P_t^2) dt} \approx \left(\frac{1}{\sum_{i=1}^n (1 - p_i)^2 \Delta_n} \right) \left(\sum_{i=1}^n \frac{(1 - p_i)(p_i - p_{i-1})}{p_i} \right). \quad (19)$$

3.2. Discrete time observations

In this case, to estimate the parameters, we suppose that the available data is the set of observations of a realization of the process at times $0 = s_0 < s_1, \dots < s_{n-1} < s_n = T$, where $\Delta_n = T/n = s_i - s_{i-1}$ ($i = 1, \dots, n$), and n is large enough. Thus, we have the discrete observations X_{s_i} from a continuous process X . We can consider the available data set as an incomplete observation of a complete data set given by the full continuous path. Then, we use diffusion bridges to complete the information and EM algorithm to find the estimators in each model. To do so, we should calculate the conditional expectation of the corresponding likelihood function for the model given the observations. Then, we need to simulate paths of the diffusion process given the data, i. e., simulate diffusion bridges.

A diffusion bridge from state a at time t_1 to state b at time t_2 ((a, t_1, b, t_2) - bridge) is a solution $\{X_t\}_{t=t_1}^{t_2}$ of a SDE such that $X_{t_1} = a$ and $X_{t_2} = b$.

To calculate the conditional expectation in E-step of EM algorithm, we use the previous estimator θ_k and we generate a diffusion bridge of size L between each couple of data, i.e.,

$$(X_{t_{i-1}}, t_{i-1}, X_{t_i}, t_i) \approx \{X_{t_{i-1}} = x_{t_{i0}}, x_{t_{i1}} \dots, x_{t_{iL-1}}, X_{t_i} = x_{t_{iL}}\},$$

where $\Delta_L = t_{i(l-1)} - t_{il} = \frac{\Delta}{L}$ for $l = 1, \dots, L$ and $i = 1, \dots, n$ with L sufficiently large such that Δ_L is close to zero.

3.2.1. Gompertz model

We use the expressions (11) and (12) in an EM algorithm to estimate MLEs corresponding. The EM algorithm works as follows. Let $\theta_0 = (b_0, \sigma_0)$ be initial values of the parameters.

Algorithm 1 EM for the Gompertz model

- (1) Set $k = 0$.
- (2) **E-step.** Calculate $\mathbb{E}_{\theta_k}[Y_t | Y_{t_{i-1}}, Y_{t_i}]$ for $t \in [t_{i-1}, t_i]$ for $i = 1, \dots, n$.
- (3) **M-step.** Using the diffusion bridges $\{(Y_{t_{i-1}}, t_{i-1}, X_{t_i}, t_i)\}_{i=1}^n$ of E-step to calculate the constants (13) we make

$$\begin{aligned} b_{k+1} &= -\frac{c_1}{\Delta_n}, \\ \sigma_{k+1} &= \sqrt{\frac{2c_3 h b_{k+1}}{1 - e^{-2b_{k+1}}}}. \end{aligned}$$

- (4) $k=k+1$ and go to 2.
-

Algorithm 1 runs K iterations with a suitable burn-in of $K_0 < K$ and then the estimators are given by

$$\hat{b}_{ML} = \frac{\sum_{k=K_0}^K b_k}{K - K_0}, \quad \text{and} \quad \hat{\sigma}_{ML} = \frac{\sum_{k=K_0}^K \sigma_k}{K - K_0}. \quad (20)$$

3.2.2. Von Bertalanffy model

Following the idea of previous sections for the case when the observations of paths are recorded at discrete times, we use expressions (15), (16) and the EM algorithm to calculate the MLEs. Let $\theta_0 = (\kappa_0, \sigma_0)$ be the initial values of the parameters, the EM algorithm corresponding to this model is as follows.

Algorithm 2 EM for the Von Bertalanffy model

- (1) Set $k = 0$.
- (2) **E-step.** Calculate $\mathbb{E}_{\theta_k}[G_t | G_{t_{i-1}}, G_{t_i}]$ for $t \in [t_{i-1}, t_i]$ for $i = 1, \dots, n$.
- (3) **M-step.** Using the diffusion bridges $\{(G_{t_{i-1}}, t_{i-1}, G_{t_i}, t_i)\}_{i=1}^n$ of E-step to calculate the constants (17) and update the estimators

$$\begin{aligned} \kappa_{k+1} &= \frac{b - a}{T} - \frac{\hat{\sigma}^2}{2}, \\ \sigma_{k+1} &= \sqrt{\frac{-a^2 + 2ab - b^2 + cn - 2dn + en}{nT}}. \end{aligned}$$

- (4) $k=k+1$ and go to 2.
-

Algorithm 2 runs K iterations with a suitable burn-in of $K_0 < K$ and then the estimators are given by

$$\hat{\kappa}_{ML} = \frac{\sum_{k=K_0}^K \kappa_k}{K - K_0}, \quad \text{and} \quad \hat{\sigma}_{ML} = \frac{\sum_{k=K_0}^K \sigma_k}{K - K_0}. \quad (21)$$

3.2.3. Logistic model

We follow [9] to estimate σ , and as in the previous models, r is estimated using diffusion bridges to complete information. The expressions (18) and (19) are used in the following EM algorithm.

Algorithm 3 Estimation of θ

- (1) Set $k = 0$.
- (2) **E-step.** Calculate $\mathbb{E}_{\theta_k}[P_t \mid P_{t_{i-1}}, P_{t_i}]$ for $t \in [t_{i-1}, t_i]$.
- (3) **M-step.**

$$r_{k+1} = \frac{\sum_{i=1}^n s_i}{n}.$$

- (4) **Update σ**

$$\sigma_{k+1} = \frac{2 \sum_{j=1}^n \sum_{l=1}^L [p_i(t_{lj}) - p_i(t_{(l-1)j})]^2}{\sum_{j=1}^n \sum_{l=1}^L [p_i(t_{lj})^2 + p_i(t_{(l-1)j})^2] \Delta_L}.$$

- (5) $k=k+1$ and go to 2.
-

To calculate the conditional expectation in E-step of Algorithm 3 we use the current θ_k and we generate a diffusion bridge to calculate s_i 's

$$s_i = \frac{1}{\sum_{j=1}^n \sum_{l=1}^L (1 - p_i(t_{l(j-1)}))^2 \Delta_L} \times \sum_{j=1}^n \sum_{l=1}^L \frac{(1 - p_i(t_{l(j-1)}))}{p_i(t_{l(j-1)})} [p_i(t_{lj}) - p_i(t_{(l-1)j})].$$

To update σ in Step 4, we use the continuous paths generated by the diffusion bridges for the E-step. Steps 2-5 are repeated until the convergence.

3.3. One record for each path

For this scenario, we will assume that each trajectory comes from the same SDE. Algorithm 4 is a new version of the method proposed in [9] to generate a path observed at times $\{t_0, t_1, \dots, t_n\}$ when we have only one measurement from each of a suitable number of M paths of the solution of the same SDE. Known θ , the parameter of SDE (1), and the constants $\alpha, \beta > 0$ to generate a path $\{X_{t_i}\}_{i=0}^n$ of the data set $\{\mathbf{X}^1, \dots, \mathbf{X}^M\}$ ($\mathbf{X}^m = \{X_{t_0}^m, X_{t_1}^m, \dots, X_{t_n}^m\}, m = 1, \dots, M$) the algorithm works as follows.

Algorithm 4 Sample one record

- (1) Draw $X_{t_0}^m$ from a $\text{Beta}(\alpha, \beta)$ distribution for $m = 1, 2, \dots, M$.
 - (2) Create a partition of the time interval $[0, T]$ into n subintervals of length Δ_n , here we choose n large enough such that Δ_n goes to zero.
 - (3) Using the Milstein scheme in the partition of the last step, draw M paths for a process that is a solution of Equation (1) with parameter θ in the time interval $[0, T]$. Let $X_{t_i}^m$ be the i th point of the m th path for $i = 0, 1, \dots, n$ and $m = 1, 2, \dots, M$.
 - (4) Make $k = 1$ and randomly choose X_{t_0} from $\{X_{t_0}^1, \dots, X_{t_0}^M\}$.
 - (5) Calculate the variance v_k of $\{X_{t_k}^1, \dots, X_{t_k}^M\}$ and drawn $Y_k \sim TN_{(X_{t_{k-1}} - v_k, \infty)}(X_{t_{k-1}}, v_k)$.
 - (6) Make $X_{t_k} = \inf_i \{X_{t_k}^i \mid X_{t_k}^i \geq Y_k\}$.
 - (7) If $k = n$, stop and the path is $\{X_{t_0}, X_{t_1}, \dots, X_{t_n}\}$. Otherwise, $k = k + 1$ and go to Step 5.
-

In Step (5) of Algorithm 4, $TN_{(a,b)}(\mu, \gamma)$ denotes a Truncated Normal random variable in (a, b) with mean μ and variance γ . In this case, we use Algorithm 4 to generate paths and we can use methods proposed in the sections 3.1 and 3.2 to estimate θ in the three models studied in this work.

4. Simulation Study

This section is devoted to presenting the results of a simulation study. We apply the methods developed in 3 to simulated data. The purpose of this is to calibrate the estimation.

4.1. Continuous observation

In this subsection, we consider the unreal case where the data consists of path observations at a continuous time for every of the three diffusion processes considered in this work. Using Milstein Scheme, we generate paths of the diffusion process with given parameters; afterward, based on these paths, we obtain the corresponding estimators of the parameters. In particular, we show numerically the consistency of our estimators.

4.1.1. Gompertz model

To illustrate the consistency of the Gompertz model estimators, we first generate paths with different time horizons. Indeed, we generate a path in the interval of time $[0, 10]$ with a discretization of $\Delta_n = 0.001$ and $n = 10,000$. We calculated the estimators for every 100 observations. The path was generated with $x_0 = 0.001$, $b = 0.6$ and $\sigma = 0.1$. In Figure 1 are plotted the estimators, which were obtained using time horizons $[0, t_k]$ for $t_k = \Delta_n k$, $k = 100, 200, \dots, 10,000$.

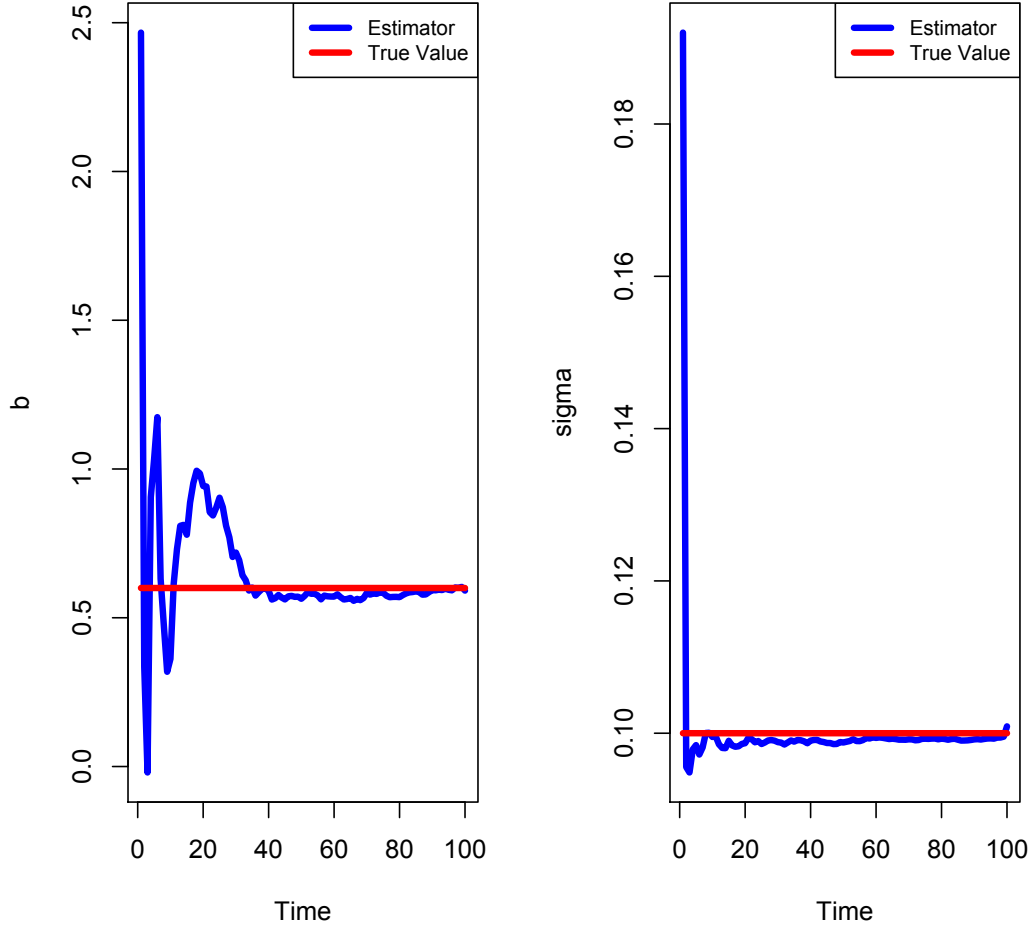


Figure 1. Consistency of estimators b and σ for the Gompertz model.

On the other hand, we simulated data set of 1,000 paths with the same parameters x_0, b, σ, Δ_n , and n in the time interval $[0, 10]$. The average and quantiles (95%) of the parameter estimator are presented in Table 1.

Table 1. Average (estimator) and quantiles (95%) of parameter estimates for Gompertz model obtained from 1,000 simulated datasets and 10,000 length of each path in the interval time $[0, 10]$.

Parameter	Real value	Estimator	Quantile 95%
b	0.6	0.59874	(0.58007, 0.61746)
σ	0.1	0.09989	(0.09857, 0.10126)

4.1.2. Von Bertalanffy model

In Figure 2 we show the asymptotic consistency of MLEs for Von Bertalanffy SDE. To create this figure, we generate paths with different time horizons $[0, t_k]$ ($k = 1, \dots, 100$) with a discretization of $\Delta_n = 0.001$ and $n = 10,000$. We calculated the estimators every 100 observations. The path was generated with $x_0 = 0.001$, $\kappa = 0.6$ and $\sigma = 0.1$.

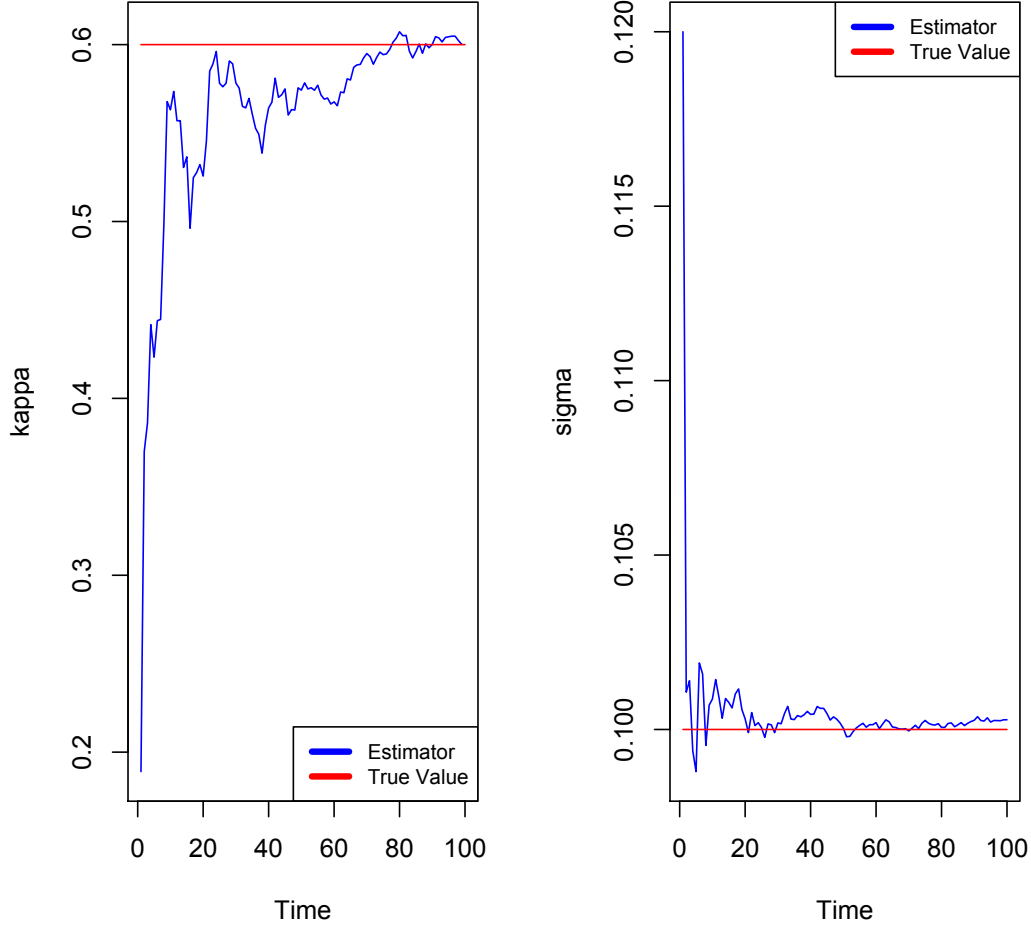


Figure 2. Consistency of estimators b and σ for the Von Bertalanffy model.

Table 2 reports the average and quantiles (95%) of parameter estimators obtained from a simulated dataset of 1,000 trajectories with parameters $x_0 = 0.001$, $\kappa = 0.6$, $\sigma = 0.1$ and $\Delta = 0.00$ in the time interval $[0, 100]$.

4.1.3. Logistic model

For this model, the chosen value for the drift parameter is $r = 0.6$. We generate the paths of the SDE (8) using the same methods that were described for the other two models. The consistency of r and σ are plotted in Figure 3.

Table 2. Average (estimator) and quantiles (95%) of parameter estimates for the Von Bertalanffy model, obtained from 1,000 simulated datasets and 10,000 lengths of each path in the interval time $[0, 100]$.

Parameter	Real value	Estimator	Quantile 95%
r	0.6	0.60086	(0.54117, 0.64894)
σ	0.1	0.10017	(0.098288, 0.10213)

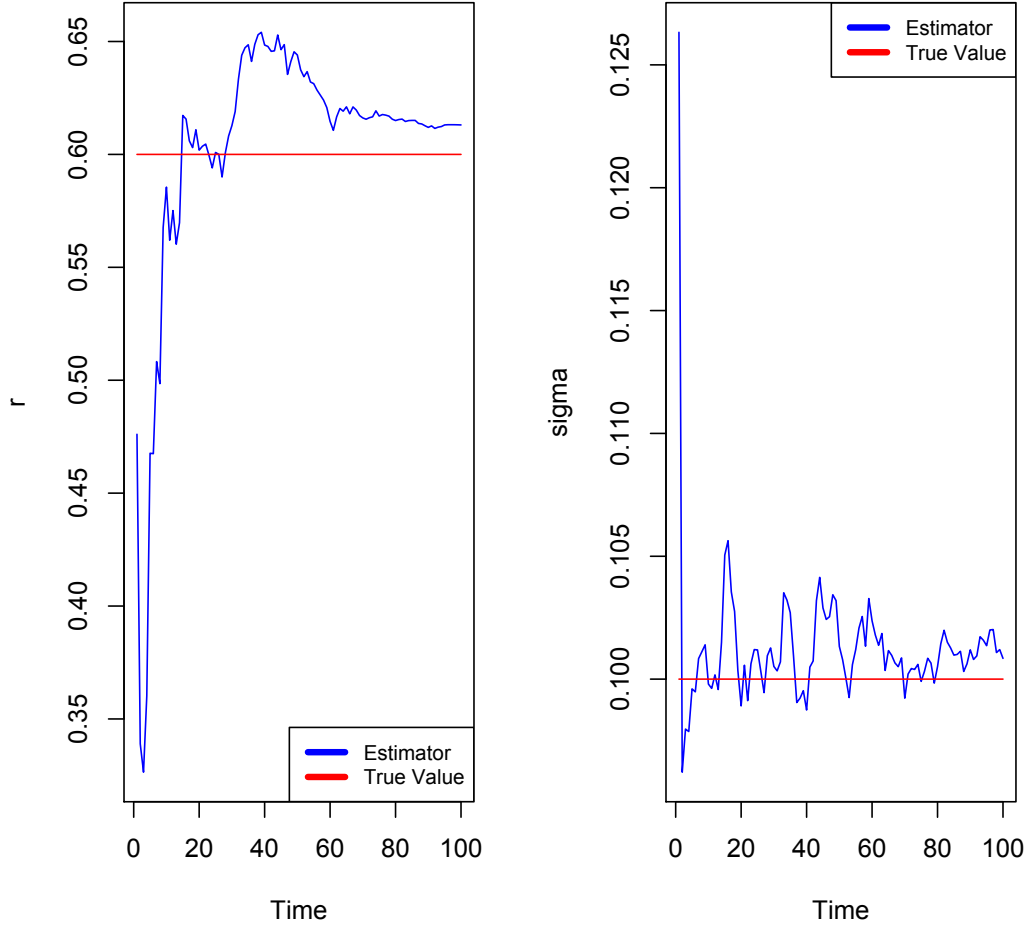


Figure 3. Consistency of estimators r and σ for the Logistic model.

Table 3 reports the average and quantiles (95%) of the corresponding estimators.

4.2. Discrete observations

In this subsection, we present the validation of our methods assuming discrete observation for the three models. For this study, 1,000 data sets of each model were simulated. Each data set

Table 3. Average (estimator) and quantiles (95%) of parameter estimates for the Logistic model, obtained from 1,000 simulated datasets and 10,000 lengths of each path in the interval time $[0, 100]$.

Parameter	Real value	Estimator	Quantile 95%
r	0.6	0.59405	(0.54464,0.64755)
σ	0.1	0.10006	(0.98728,0.10151)

was obtained by simulating a sample path of length 10,001 in the interval time $[0, 100]$ with an initial value $x_0 = 0.001$. We suppose that we have observed only 1,001 points of the path at times $0 = t_0, 1 = t_1, \dots, t_{1000} = 100$ ($n = 1001$). This means $\Delta_n = 0.1$ for each path.

4.2.1. Gompertz model

In this subsection, we present the result of a simulation study for the Gompertz model that satisfies the SDE (2). The given parameter values were $b = 0.6$ and $\sigma = 0.1$. We run Algorithm 1 with the initial values $b_0 = 0.7$ and $\sigma_0 = 0.21$ and $L = 1,000$. Figure 4 plots the estimators of 1,000 iterations of Algorithm 1 for a path of the solution of SDE (2).

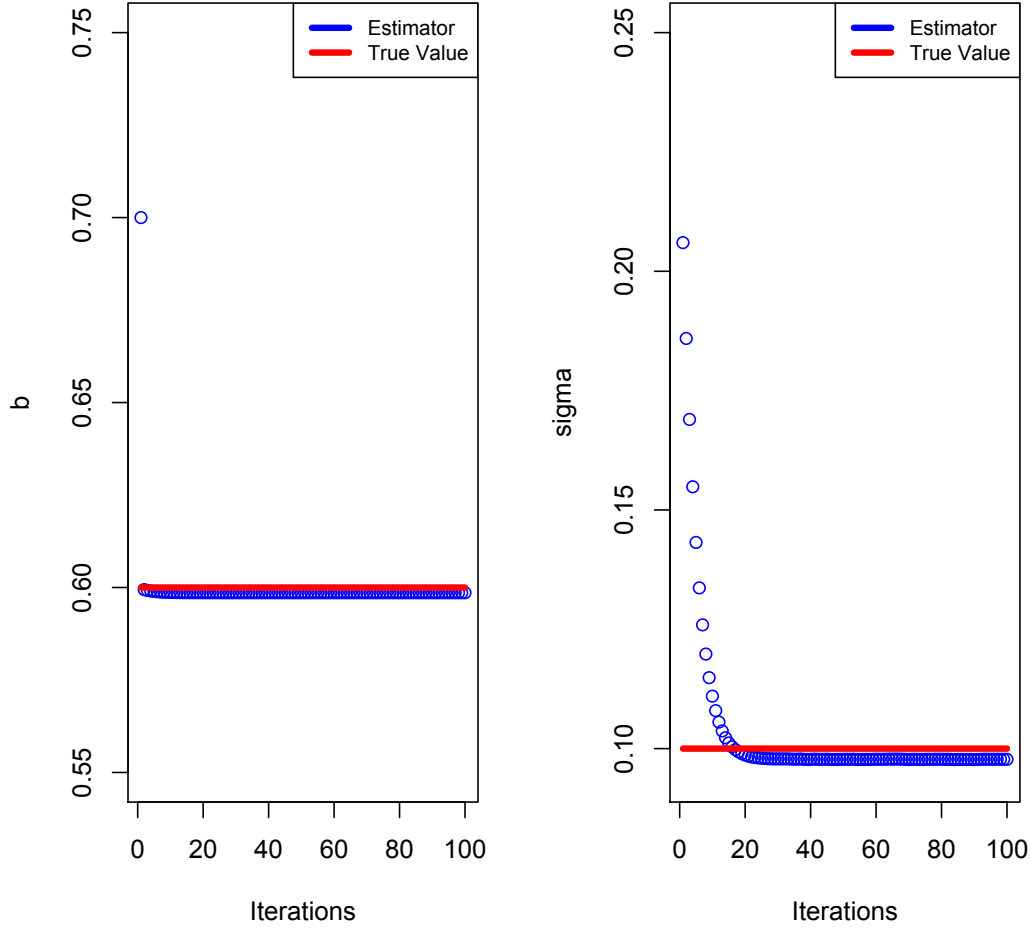


Figure 4. Iterations of EM algorithm to estimate the parameters b and σ for the stochastic Gompertz model.

The average and quantiles (95%) of parameter estimators obtained from the sample at the 100th iteration are presented in Table 4.

Table 4. Average (estimator) and quantiles (95%) of parameters for the Gompertz model.

Parameter	Real value	Estimator	Quantile 95%
b	0.6	0.59943	(0.59652,0.60286)
σ	0.1	0.10059	(0.09958,0.10103)

4.2.2. Von Bertalanffy model

We now present the corresponding results for the Von Bertalanffy model. For this case, the parameter values were fixed to $\kappa = 0.6$ and $\sigma = 0.1$. Figure 5 shows 100 iterations of Algorithm 2 with $\kappa_0 = 0.2$, $\sigma_0 = 0.15$ and $L = 1,000$.

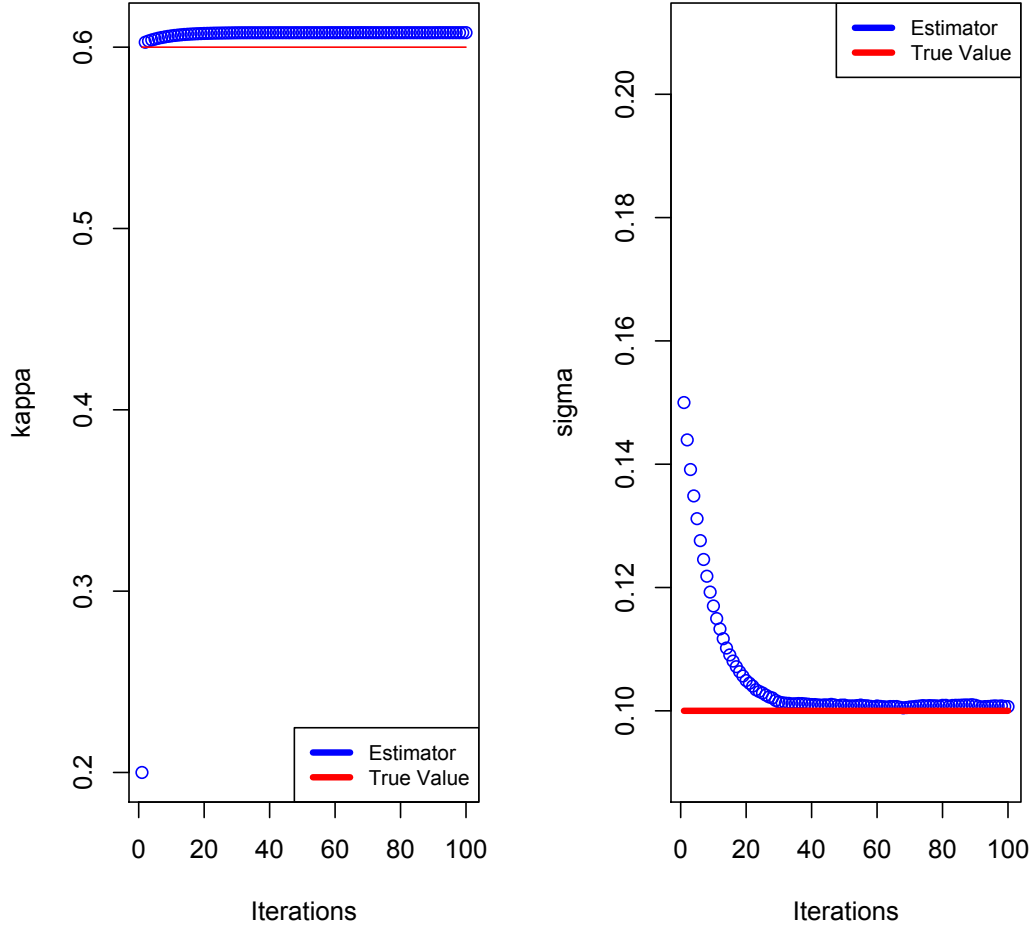


Figure 5. Iterations of EM algorithm 2 to estimate the parameters κ and σ for the stochastic Von Bertalanffy model.

Table 5 reports the average and the quantiles (95%) of the last iterations of the data sets of 1,000 paths observed at discrete time.

Table 5. Average (estimator) and quantiles (95%) of parameters κ and σ in the Von Bertalanffy model.

Parameter	Real value	Estimator	Quantile 95%
κ	0.6	0.59511	(0.57946,0.60992)
σ	0.1	0.10381	(0.09938,0.10761)

4.2.3. Logistic model

This subsection is devoted to showing the results of a simulation study for the Logistic stochastic model when we have observations at discrete times. We present results with parameters

values of $r = 0.6$ and $\sigma = 0.1$. First, we illustrate Algorithm 3 in Figure 6 with 100 iterations, $r_0 = 0.5, \sigma_0 = 0.2$ and $L = 1,000$.

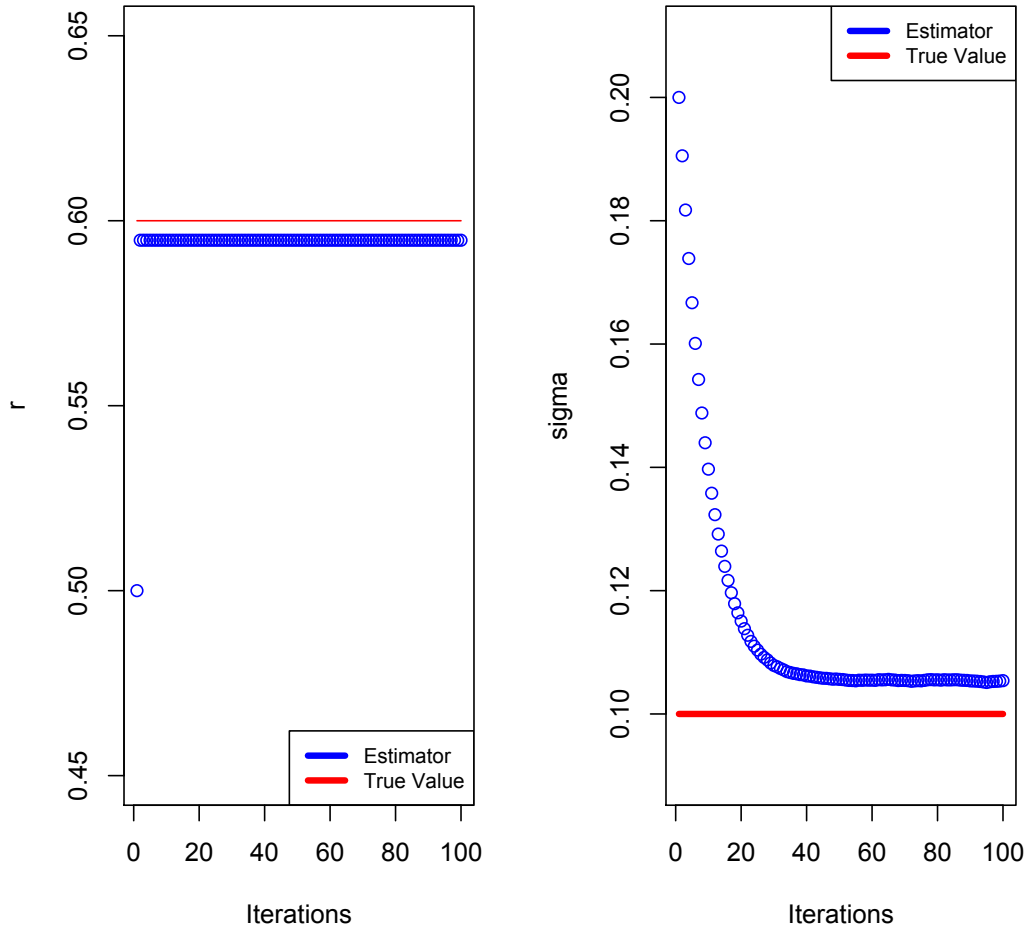


Figure 6. Iterations of EM algorithm 3 to estimate the parameters r and σ for the stochastic Logistic model

The average and the quantiles (95%) corresponding are presented in Table 6.

Table 6. Average (estimator) and quantiles (95%) of parameter estimates for the Logistic model, obtained from a sample of size 1,000 of the 100 th iteration of Algorithm 3

Parameter	Real value	Estimator	Quantile 95%
r	0.6	0.58324	(0.52946,0.61549)
σ	0.1	0.10923	(0.09323,0.11314)

4.3. Validation of the method assuming one record

In this section we present a simulation study for the extreme case, in which we get only one observation of different paths; here, however, we assume we have observed several paths from the same SDE. We assumed they are observed at discrete times.

4.3.1. Gompertz model

We used Algorithm 4 to generate a path of a process solution of the equation (2). Algorithm was run for $M = 100$, $n = 10,000$ and $\Delta_n = 0.001$. We fixed the parameter, to generate each path, as $b = 0.6$ and $\sigma = 0.1$ with initial distribution $Beta(1, 100)$. We suppose that we have only 1001 observations at times $0 = t_0, 1 = t_1, \dots, t_{1000} = 1$ ($k = 1001$). Then $\Delta_k = 0.01$ for each path. Thus, using Algorithm 4 we have a trajectory observed at discrete time and we can use Algorithm 1 to find the corresponding MLEs. We run Algorithm 1 with 50 iterations and initial values $b_0 = 0.5$ and $\sigma_0 = 0.2$. Figure 7 shows the iterations of Algorithm 1.

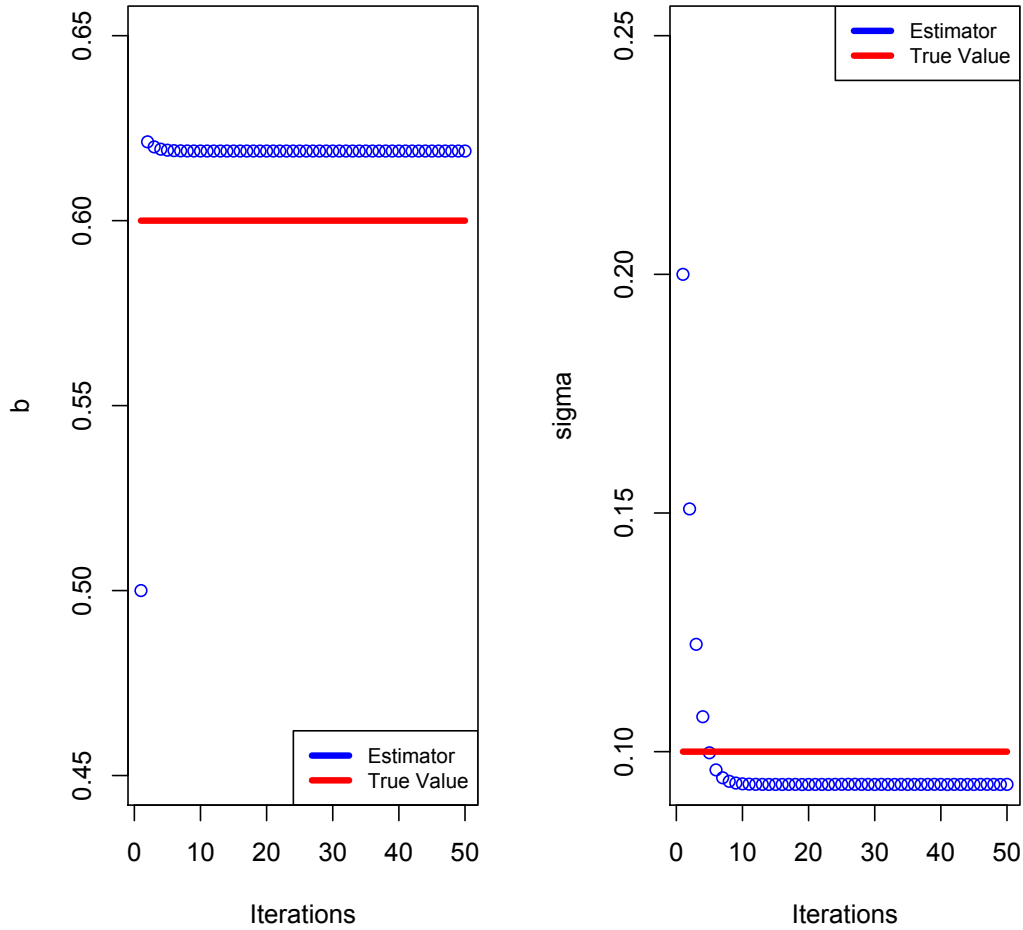


Figure 7. Iterations of EM algorithm 1 to estimate the parameters b and σ for the stochastic Gompertz model.

4.3.2. Von Bertalanffy model

We now apply Algorithm 4 to generate a path observed at discrete time in Von Bertalanffy model where the real values of parameters are $\kappa = 0.6$ and $\sigma = 0.1$ and the initial value follows a $Beta(1, 100)$. We run Algorithm 4 with $M = 100$, $n = 10,000$ and $\Delta_n = 0.001$ where we have only 1001 observations at times $0 = t_0, 1 = t_1, \dots, t_{1000} = 1$ ($k = 1001$). Using the path generate by Algorithm 4 we apply Algorithm 2 to find the corresponding MLEs. Algorithm 2 was run with 50 iterations and initial values $\kappa_0 = 0.4$ and $\sigma_0 = 0.25$. In Figure 8 we can observe the evolution of Algorithm 2.

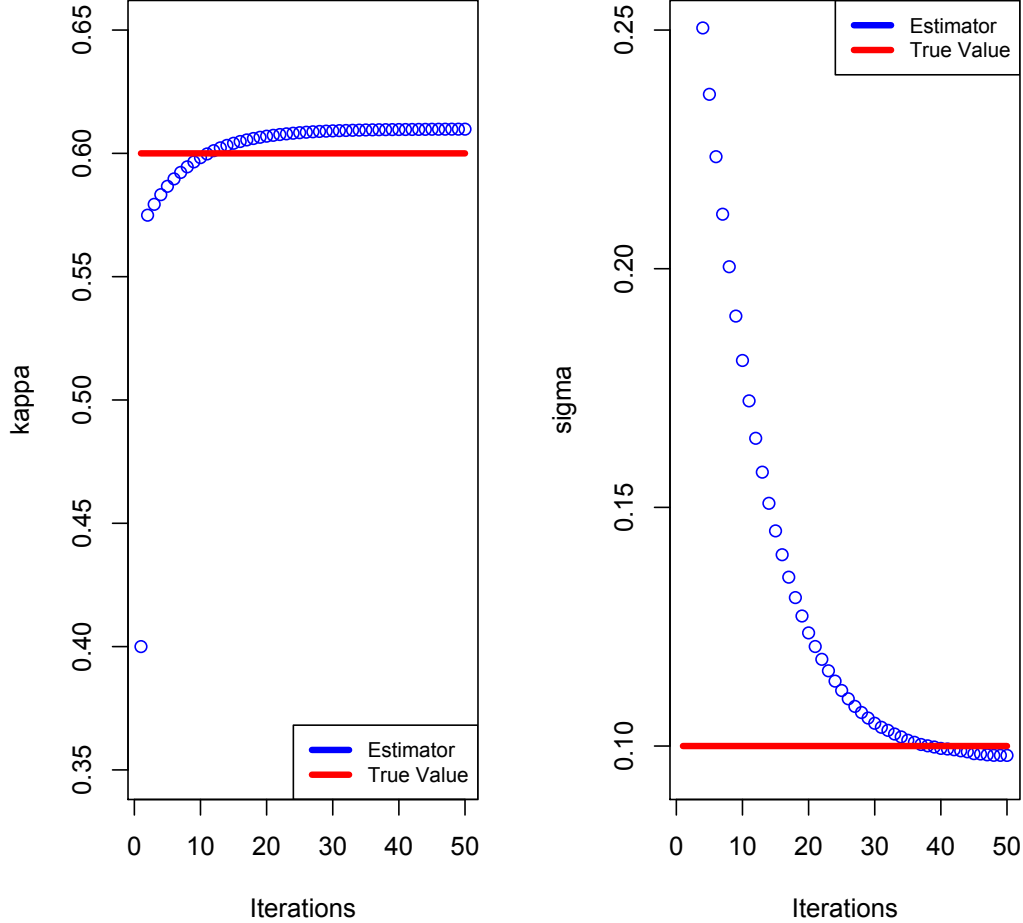


Figure 8. Iterations of EM algorithm 2 to estimate the parameters κ and σ for the stochastic Von Bertalanffy model.

4.3.3. Logistic model

Finally, we make the same simulation study for the Logistic model. The corresponding parameters to generate data are $r = 0.6$ and $\sigma = 0.1$. The other parameters of Algorithm 4 are the same that we used in the two sections previous. Algorithm 3 was run with 50 iterations and initial values $r_0 = 0.8$ and $\sigma_0 = 0.3$. Figure 9 shows the obtained results.

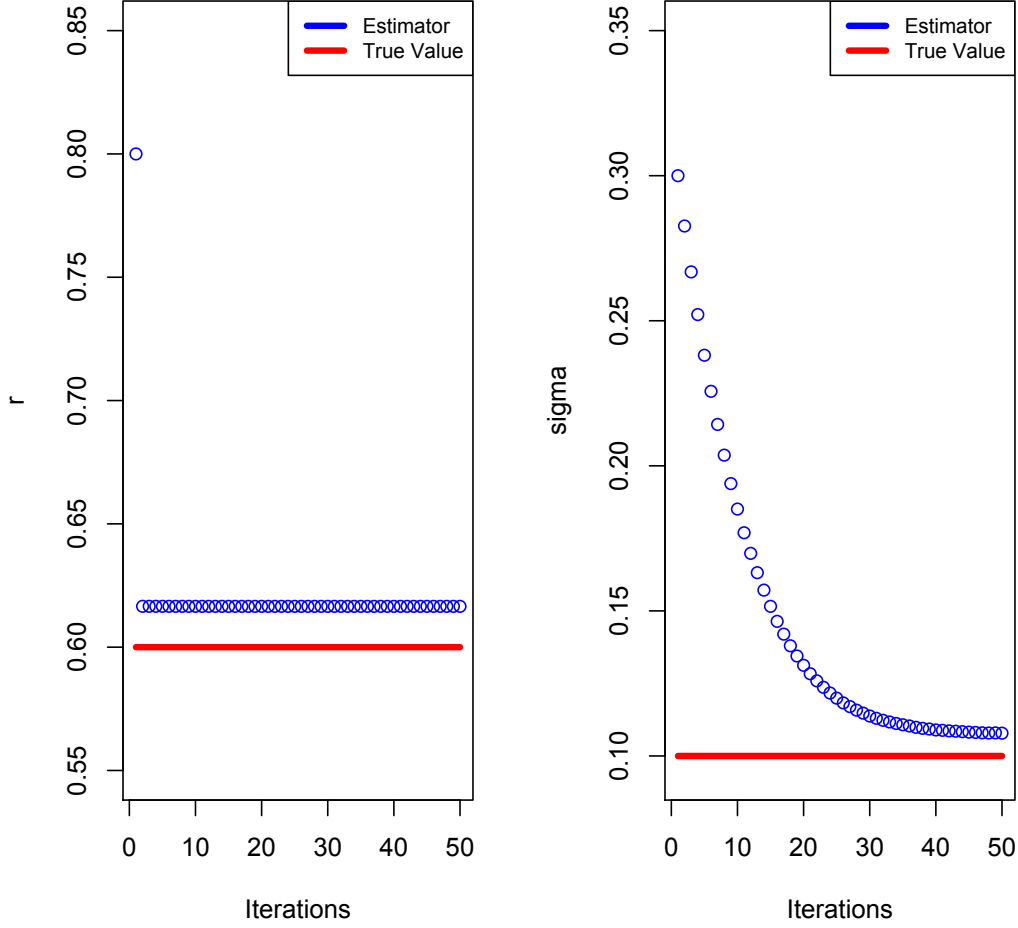


Figure 9. Iterations of EM algorithm 3 to estimate the parameters r and σ for the stochastic Logistic model.

5. Model selection via AIC

5.1. Estimators for each path and model.

Akaike's information criterion, referred to as AIC, is a method for evaluating and comparing statistical models developed by the Japanese mathematician Hirotugu Akaike. It provides a measure of the quality of the estimate of a statistical model taking into account both the goodness of fit and the complexity of the model. [5]. The well-known Akaike information criterion is given by

$$\text{AIC} = -2 \log L(\hat{\theta} \mid \text{data}) + 2\kappa_{\theta}, \quad (22)$$

where $L(\hat{\theta} \mid \text{data})$ is the likelihood, $\hat{\theta}$ is an estimator for the parameter (or parameters) and κ_{θ} is the total number of parameters in the model. Usually, $\hat{\theta}$ is taken as the maximum likelihood estimator $\hat{\theta}_{MLE}$.

Consider the two stochastic differential equations:

$$\begin{aligned} dX_t &= \alpha F(X_t)dt + \sigma G(X_t)dB_t, \\ X_0 &= x_0, \end{aligned} \tag{23}$$

and

$$\begin{aligned} dX_t &= \sigma G(X_t)dB_t, \\ X_0 &= x_0. \end{aligned} \tag{24}$$

Denote by \mathbb{P}_α and \mathbb{P} the probability measures induced by each solution of the equations, respectively.

Thus, as was mentioned before, \mathbb{P}_α and \mathbb{P} are equivalent (see [12]) and the corresponding Radon-Nikodym derivative, which means the Likelihood, is given by the Girsanov theorem

$$\frac{d\mathbb{P}_\alpha}{d\mathbb{P}}(X) = \exp \left[\int_0^T -\alpha \frac{F(X_t)}{\sigma^2 G^2(X_t)} dX_t - \frac{1}{2} \int_0^T \alpha^2 \frac{F^2(X_t)}{\sigma^2 G^2(X_t)} dt \right]. \tag{25}$$

Then, the log-likelihood is defined as

$$\log L(\alpha) = \int_0^T -\alpha \frac{F(X_t)}{\sigma^2 G^2(X_t)} dX_t - \frac{1}{2} \int_0^T \alpha^2 \frac{F^2(X_t)}{\sigma^2 G^2(X_t)} dt. \tag{26}$$

Thus, the AIC for the stochastic model (23) is

$$\text{AIC} = -2 \log L(\hat{\alpha}_{ML}) + 2k, \tag{27}$$

with k being the number of estimated parameters in the model.

Remark 5.1. *To obtain the AIC for the stochastic Gompertz model we apply (26) with $F(x) = x \log(x)$ and $G(x) = x$.*

The AIC for the stochastic Von Bertalanffy model we apply (26) to the functions $F(x) = G(x) = L_\infty - x$.

Finally, the AIC for the stochastic Logistic model we use (26) with the functions $F(x) = x(1-x)$ and $G(x) = x$.

5.2. A numerical example

This section is devoted to studying a numerical example of the AIC implementation for the three SDEs. We proceed as follows. First, we generate a sample path for each SDE studied in this work. Afterward, we use our methods to estimate the corresponding parameter for each SDE. At this point, we assume we do not know what is the "true" model, then we fit every sample path for the three SDEs. This implies that we have for every sample path three possible models and to choose the best model we apply the AIC criteria. To do that, we calculate the AIC for each case and verify that the "true" model is the one that minimizes the AIC.

We generate the three paths using the same initial value $x_0 = 0$, for Gompertz model the parameter $b = 0.6$, for Von Bertalanffy $\kappa = 0.6$ and $r = 0.6$ in Logistic case. The value of the parameter $\sigma = 0.1$ for all models and we generate the paths in the time interval $[0, 10]$ with a discretization of $\Delta_n = 10/n$ where $n = 10,000$. Figure 10 shows paths for all models using corresponding parameters.

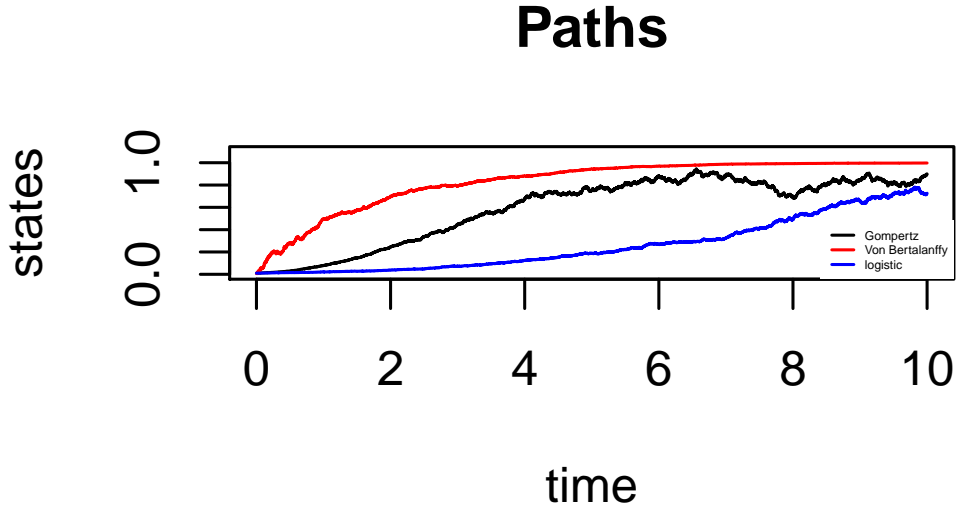


Figure 10. Paths

Table 7 reports the value of the parameters of every model with the description given above. We use these parameters to calculate the AIC. We observe that we are assuming that the log-likelihood is a function only of the drift parameters.

Table 7. Estimators fitted for each model.

Parameter	True Model	Method	Real value	Estimator
b	Gompertz	Gompertz	0.6	0.60443
σ	Gompertz	Gompertz	0.1	0.09982
b	Von Bertalanffy	Gompertz	0.6	3.91481
σ	Von Bertalanffy	Gompertz	0.1	0.47852
b	Logistic	Gompertz	0.6	0.14383
σ	Logistic	Gompertz	0.1	0.09904
κ	Gompertz	Von Bertalanffy	0.6	0.22904
σ	Gompertz	Von Bertalanffy	0.1	0.13125
κ	Von Bertalanffy	Von Bertalanffy	0.6	0.59583
σ	Von Bertalanffy	Von Bertalanffy	0.1	0.105623
κ	Logistic	Von Bertalanffy	0.6	0.12857
σ	Logistic	Von Bertalanffy	0.1	0.04493
r	Gompertz	Logistic	0.6	1.41950
σ	Gompertz	Logistic	0.1	0.100101
r	Von Bertalanffy	Logistic	0.6	4.39651
σ	Von Bertalanffy	Logistic	0.1	0.03536
r	Logistic	Logistic	0.6	0.590984
σ	Logistic	Logistic	0.1	0.10030

In Table 8 are reported the AIC given a true model and fitting the three models to each set of the simulated data. To illustrate the table if we take the Gompertz SDE as the true model, then the AIC when it is assumed the Gompertz model is -919.09, for the Von Bertalanffy

is 301.39, and for the Logistic is -283.42. Thus, from the results contained in this table, we conclude that in the three cases the AIC criterion allows us to choose the "true" model.

Table 8. AIC for each model and method.

Fit Model - True Model	Gompertz	Von Bertalanffy	Logistic
Gompertz	-919.09	-30.44	1534.14
Von Bertalanffy	301.39	-318.18	38389.18
Logistic	-283.42	-81.88	628.95

6. Concluding remarks

In this paper, we have presented a method to fit SDEs to different types of datasets as follows. To illustrate the model, we consider as examples three stochastic models for biological growth. Each model is given by different stochastic differential equations which are a stochastic version of classical deterministic differential equations, which have been applied to several fields of sciences, and therefore they are very important. We have considered SDEs driven by an affine noise.

We assumed that fitting the stochastic model is equivalent to adjusting only a fixed number of parameters in each SDE. Furthermore, we considered that these parameters are constants but unknown and we estimated them. We have shown a method to estimate these parameters for different types of datasets. Indeed, the dataset we successfully managed could be a path of the SDE that have discrete or continuous observations. Moreover, the dataset could have only one observation for each path, subject to the condition that we have a sufficient number of observed paths. This permits us to rebuild (continuous or discrete) paths to estimate the parameters.

Finally, in the method presented here, we have used the classical Akaike information criterion (AIC) to select the best model. We have used Girsanov's Theorem to define the log-likelihood and thus the AIC. We have run simulations to validate the estimation procedure and the selection of the best model. For the simulated data, we have found that the AIC provides, as in the deterministic case, a good tool for selecting models.

7. Acknowledgements

Baltazar-Larios F. has been supported by UNAM-DGAPA-PASPA.

8. Declarations

Conflict of interest The authors declare that they have no competing interests.

References

- [1] Ando, T. (2010). *Bayesian model selection and statistical modeling*. North Carolina State University, North Carolina State University.
- [2] Bladt, M. and Sørensen, M. (2014). Simple simulation of diffusion bridges with application to likelihood inference for diffusions. *Bernoulli*, 20(2):645–675.
- [3] Braun, M., Coleman, C. S., Drew, D. A., and Lucas, W. F. (1983). *Differential equation models*, volume 1. Springer-Verlag, New York.
- [4] Braun, M. and Golubitsky, M. (1992). *Differential equations and their applications*, volume 1. Springer-Verlag, New York, 4 edition.

- [5] Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference*. A practical information-theoretic approach. Springer-Verlag, New York.
- [6] Celeux, G. and Diebolt, J. (1986). The sem algorithm: A probabilistic teacher algorithm derived from the em algorithm for mixture problem. *Comput. Statist*, pages 599–613.
- [7] Chao, W. and Huisheng, S. (2016). Maximum likelihood estimation for the drift parameter in diffusion processes. *Stochastics*, 88(5):699–710.
- [8] DeAngelis, D. L. and Gross, J. (2018). *Individual-based models and approaches in ecology: populations, communities and ecosystems*. CRC Press, United States.
Scandinavian Journal of Statistics, 40(2):322–343.
- [9] Delgado-Vences, F., Baltazar-Larios, F., Ornelas-Vargas, A., Morales-Bojórquez, E., Cruz-Escalona, V. H., and Salomón Aguilar, C. (2022). Inference for a discretized stochastic logistic differential equation and its application to biological growth. *Journal of Applied Statistics*.
- [10] Dempster, A., Laird, N., and D.B., R. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *J. R. Stat. Soc., Ser. B Stat. Methodol*, 39:1–38.
- [11] Hasanoglu, A. H. and Romanov, V. G. (2017). *Introduction to inverse problems for differential equations*. Springer International Publishing, New York.
- [12] Iacus, S. M. (2009). *Simulation and inference for stochastic differential equations: with R examples*. Springer Science & Business Media, New York.
- [13] Isakov, V. (2006). *Inverse problems for partial differential equations*, volume 127. Springer, New York.
- [14] Jiang, D. and Shi, N. (2005). A note on nonautonomous logistic equation with random perturbation. *Journal of Mathematical Analysis and Applications*, 303(1):164–172.
- [15] Lillacci, G. and Khammash, M. (2010). Parameter estimation and model selection in computational biology. *PLOS Computational Biology*, 6(3):1–17.
- [16] Moss, D. K.; Ivany, L. C.; Jones, D. S. (2021). Fossil bivalves and the sclerochronological reawakening. *Paleobiology*: 1–23
- [17] Nielsen, S. F. . (2000). The stochastic em algorithm: estimation and asymptotic results. *Bernoulli*, 6(3):457–489.
- [18] Paek, J. and Choi, I. (2014). Bayesian inference of the stochastic gompertz growth model for tumor growth. *Communications for Statistical Applications and Methods*, 21(6):521–528.
- [19] Panik, M. J. (2017). *Stochastic Differential Equations: An Introduction with Applications in Population Dynamics Modeling*. John Wiley & Sons, United Kingdom.
- [20] Pavliotis, G. A. (2014). *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, New York.
- [21] Román-Román, P., Romero, D., and Torres-Ruiz, F. (2010). A diffusion process to model generalized von bertalanffy growth patterns: Fitting to real data. *Journal of Theoretical Biology*, 263(1):59–69.
- [22] Wei-Cheng, M. (2006). Estimation of diffusion parameters in diffusion processes and their asymptotic normality. *Int. J. Contemp. Math. Sciences*, 1(16):763 – 776.