Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica D 213 (2006) 190–196
PHYSICA D
www.elsevier.com/locate/physd

Selecting nonlinear stochastic process rate models using information criteria

David M. Walker[a,*], Glenn Marion[b]

[a] *Biomathematics and Statistics Scotland, The Macaulay Institute, Craigiebuckler, Aberdeen AB15 8QH, United Kingdom*
[b] *Biomathematics and Statistics Scotland, James Clerk Maxwell Building, King's Buildings, Edinburgh EH9 3JZ, United Kingdom*

Received 3 May 2005; received in revised form 7 November 2005; accepted 22 November 2005
Available online 20 December 2005
Communicated by H. Levine

Abstract

We demonstrate how unknown process rates within a stochastic modelling framework based on Markov processes can be approximated from time series data using polynomial basis functions. The problem of model selection is considered by adapting basis function selection methods and the minimum description length information criteria which have previously been developed for nonlinear autoregressive models of time series under Gaussian noise assumptions. We investigate the effectiveness of the methods with application to stochastic biological population models.

Keywords: Stochastic process models; Model selection; Description length; Nonlinear models

1. Introduction

Stochastic modelling and simulation have been widely applied to understand and describe the behaviour of a range of complex phenomena including biological populations [1–6], epidemic dynamics [7] and chemical reactions [8]. The theory of stochastic processes [9] is also a natural framework in which to study so-called agent-based models in which agents interact with each other and their environment using simple local rules. For example, in economics one can think of an agent buying, selling, or holding stock on the basis of limited information, in ecology a grazing animal may choose to graze a particular location, or decide to forage depending on the local availability of resources or individual energy requirements [10]. The stochastic approach to modelling can not only account for variability and spatial heterogeneity, but also point the way towards better deterministic representations which model such variation using suitable limiting processes and approximations [11,7,2,3,5,12,10].

Despite the widespread use of stochastic process models techniques which link them to observational data are somewhat limited and methods are needed to estimate (the distribution of) parameter values from data and to perform model selection, that is to select the model (from some defined class) which is best supported by the data. Recent advances in computational methods such as Markov chain Monte Carlo (MCMC) have enabled parameter estimation for discrete-time Markov models [13] and continuous time Markovian [14–16] and non-Markovian processes [17]. MCMC methods have also been applied to enable model selection between simple epidemic models [18]. However, although in principle very flexible, MCMC methods are computationally intensive and more worryingly, except in special cases [19], there are no general results allowing a decision to be made as to when, or if, the Markov chain has converged to the distribution of interest and therefore heuristic criteria are typically employed [20].

In this paper we avoid the problems associated with MCMC by tackling model selection in Markov process models by discriminating between competing models using a novel application of a basis selection algorithm [21] and the minimum description length (MDL) principle [22] previously developed for model selection in nonlinear time series reconstruction. In [21] a deterministic predictive model of a system in an equivalent phase space is reconstructed from time series data under Gaussian noise assumptions. Our approach differs in that we attempt to reconstruct deterministic process rates of

* Corresponding author.
E-mail addresses: d.walker@bioss.ac.uk (D.M. Walker), glenn@bioss.ac.uk (G. Marion).

0167-2789/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.physd.2005.11.007

a stochastic systems model. An essential difference manifests itself in the likelihoods being given by products of exponential distributions instead of Gaussian distributions.

We will also represent the rate functions (see Section 2) in our stochastic population models by polynomial functions and employ the MDL criteria to select the appropriate number of terms solely on the basis of time series data. We recognize that polynomials are not the most suitable or best choice of approximating function and have chosen to use them only for ease of exposition. Indeed in our first example where the perfect model is given by polynomial process rates their use helps raise a point highlighting why the best MDL-selected model is not always the true representation.

The minimum description length principle has been used with some success in discriminating between different model approximations to dynamics reconstructed from nonlinear time series. Judd and Mees [21,23] have used MDL to reconstruct radial basis function models, Small and Tse [24] have demonstrated how MDL can select between different neural network architectures, and description length has been applied to linear stochastic processes by Small and Judd [25]. Nakamura [26,27] has compared the performance of MDL with other information criteria such as Akaike's Information Criterion [28] and Schwarz's Information Criterion [29] for linear autoregressive models, polynomial basis models and radial basis function models. All of the above works involve additive Gaussian noise assumptions and so our approach widens the application of description length methods to a new model class with a different noise source.

The outline of this paper is as follows: In Section 2 we introduce the framework of stochastic process rate models and describe how for such Markov processes the event rates can be approximated with nonlinear functions. In Section 3.1 we briefly outline the subset selection method of Judd and Mees [21] and discuss the important modifications required for the class of stochastic models we study. We describe model selection based on the minimum description length principle in Section 3.2. These methods and ideas are explored in Section 4 through application to two (stochastic) population models from ecology: logistic population growth in Section 4.1, and a predator–prey population model in Section 4.2. We close the paper with a short discussion and concluding remarks in Section 5. A preliminary outline of this work has been presented elsewhere [30].

2. Stochastic models and nonlinear function approximations

In the framework of Markov process population dynamics the simple rules of agent-based models are described by a sequence of events determined probabilistically. The event probabilities share the common structure

$$P(\mathbf{s}(t+\delta t) = \mathbf{s}(t) + \delta\mathbf{s}) = r_j(\mathbf{s} \to \mathbf{s} + \delta\mathbf{s}; \lambda)\delta t \tag{1}$$

where $r_j(\mathbf{s} \to \mathbf{s} + \delta\mathbf{s}; \lambda)$ represents the (probability) rate of an event of type j which causes the change $\delta\mathbf{s}$ in state space in time interval $(t, t + \delta t)$. If there are m possible events with rates r_j, $j = 1, \ldots, m$, and the system is in state $\mathbf{s}(t)$ then the total event rate at time t is $R(\mathbf{s}(t)) = \sum_{j=1}^{m} r_j(\mathbf{s}(t))$ and the time to the next event τ is exponentially distributed $\sim e^{-R(\mathbf{s}(t))\tau}$. For example, in a population model these rates could represent the probability of *only* a birth event or a death event occurring in the short time interval.[1] The λ represent parameters of the stochastic process.

A simulated realization z of the Markov process [1] is generated from an initial state $\mathbf{s}(t_1)$ by randomly selecting the inter-event time from the exponential distribution $\tau \sim e^{-R(\mathbf{s}(t_1))\tau}$ and applying (1) to choose one of the m possible events, thereby determining $\mathbf{s}(t_1)$. Iterating this procedure generates

$$z = \{\mathbf{s}(t_1), \ldots, \mathbf{s}(t_N)\} \tag{2}$$

where $t_1, \ldots, t_N$ denote the event times and $\mathbf{s}(t)$ denotes the state of the system immediately prior to the event at time t. The model is event driven so let n index these events and denote the event type occurring at time t_n by $E(n) \in \{1, \ldots, m\}$. If all event types are visible and if the process is monitored continuously we can observe all N events of the realization, so that the data $D = z$. The likelihood $L(\boldsymbol{\lambda}, D)$ of the process may be written as

$$L(\lambda, D) = \prod_{n=1}^{N} r_{E(n)} e^{-(t_n - t_{n-1}) \sum_{j=1}^{m} r_j(\mathbf{s}(n))} \tag{3}$$

where the observations D are equivalent to the complete realization z, the state of the system for all $t \in (t_0, t_N)$. The rate $r_{E(n)}$ corresponds to the probability of event n occurring at time t_n whilst the exponential term is the probability that nothing happens between event $n-1$ and event n. When the parameter space is low dimensional, and the data is a complete realization, it is straightforward to obtain maximum likelihood estimates of the parameters by applying standard optimization routines.

We propose to represent the (unknown) rate functions in a stochastic process model by nonlinear functions. In this paper we consider polynomial basis functions but other representations are possible. We note there is no need to select event rate functions taking values between 0 and 1 as they are suitably scaled to probabilities later as in Eq. (1). Therefore, thin-plate splines, radial basis function networks, and even neural network function approximations could perhaps provide better models for a given application since polynomials may not the best class of fitting functions to choose. If the state of the system at the time of event n is denoted by $\mathbf{s}(n)$ then a polynomial approximation to a process rate is

$$R(\mathbf{s} \to \mathbf{s} + \delta\mathbf{s}) = \mathbf{A} + \mathbf{B}\mathbf{s}(n) + \mathbf{C}\mathbf{s}(n)\mathbf{s}^T(n) + \cdots \tag{4}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$ are constant tensors of appropriate rank.

[1] We are indulging in an abuse of terminology here by referring to "probability" and "rate" interchangeably. Strictly speaking (1) describes the process rates which by a simple scaling using the total event rate R can be converted to probabilities.

The values of the parameters can be estimated by minimizing the negative log likelihood where the rate functions in (3) are replaced by the appropriate polynomial representations (4). Our focus here, however, is on model selection within a chosen model class. In particular how do we determine the appropriate set of basis functions (polynomial degree) given an observed realization of the (unknown) process? A crucial problem associated with nonlinear basis functions, and polynomial approximations in particular, is the bias-variance dilemma and the curse of dimensionality; models with a large number of parameters are difficult to estimate and do not generalize well. We address this issue by appealing to the minimum description length principle [22].

3. Model estimation

The minimum description length principle is an information theoretic approach to balancing complexity and predictability for selecting between competing models explaining the data. The idea is to calculate the description length required to explain the observed data relative to some model and information required to specify the model and its parameters. The model for a given class which returns the minimum description length value is regarded as the best model describing the process given the data. For a given class of models such as the polynomial functions outlined above it is typically unfeasible to calculate the description length of all possible model representations.

To proceed it is therefore necessary to have a practical method for exploring a (large) subset of the model class. A procedure demonstrated to be successful is a subset selection method introduced by Judd and Mees [21] which we now describe in our context.

3.1. Selection algorithm

In this paper we adapt the selection algorithm developed by Judd and Mees [21] to select subsets of basis functions of nonlinear models under Gaussian noise assumptions in the context of reconstructing time series models from data. The sensitivity analysis results of optimization theory are applied to the likelihood in order to determine which model basis function should be selected for removal or inclusion.

Specifically, it gives a prescription for selecting a new basis function to add to a current model which minimizes a cost function related to the likelihood, or for selecting an existing basis function to remove from the current representation and does the least "damage" in the sense of increasing a cost function. This so-called "breathing" approach helps search for candidate models within a model class in a focused way.

This step is computationally slightly more complicated in our present situation due to the non-Gaussian likelihood but here we show that the arguments carry through in a similar manner to those presented by [21].

Following Judd and Mees, the problem of selecting which basis function to add to the current model of the stochastic rate approximation can be written as[2]

$$\text{minimize } -\log L(\lambda) \text{ subject to } N(\lambda) = k, \tag{5}$$

where we have dropped the dependence on data in the likelihood for clarity in exposition. The constraint $N(\lambda) = k$ corresponds to the number of parameters in the model. Setting $B = \{j : \lambda_j \neq 0\}$, so that $N(\lambda) = \mid B \mid$, we can use sensitivity analysis to see the effect of changing the size of B. We write the constraint as

$$\lambda_j = u_j, \quad j \notin B \tag{6}$$

where the $\mathbf{u} = \mathbf{0}$ but are kept as parameters. The Lagrangian [31] for (5) with (6) is

$$L(\lambda, \mu) = -\log L(\lambda) + \mu^T(\mathbf{u} - \lambda) \tag{7}$$

where μ are dual variables. The Kuhn–Tucker conditions for minimizing this give rise to

$$\mu = \nabla_\lambda(-\log L(\lambda))$$
$$\mathbf{u} - \lambda = 0.$$

Since μ is the dual variable corresponding to constraint (6) it is the sensitivity to changes in $\mathbf{u}$ at optimality, and therefore the largest element of μ in absolute value corresponds to the basis function that should be added to give the greatest marginal improvement in cost. This gives a prescription for adding a basis function to extend the model.

We can similarly find a prescription for removing an existing basis function from the model by considering the (Lagrangian) dual problem. That is,

$$\text{maximize } -\log L(\lambda) + \mu^T(\mathbf{u} - \lambda) \text{ subject to } \mu_j = w_j,\ j \in B$$

where the $\mathbf{w} = \mathbf{0}$ but are kept as parameters and $\lambda = \lambda(\mu)$. If we set $\mathbf{u} = \mathbf{0}$ immediately then the Kuhn–Tucker conditions show that the Lagrange multiplier for the constraints in this problem is a dual variable of λ and so selecting the smallest existing λ_j in absolute value as the variable to remove from the basis will do the least damage to the cost.

The above information gives a means by which an iterative scheme for expanding and shrinking model size can operate. Following [21] the 'best' model over all model sizes can be chosen as the one which minimizes a description length criterion to be described in the next section.

3.2. Description length

Rissanen's [22] minimum description length principle can be used to discriminate between models of different size for the same data set by comparing the cost of describing the data in

[2] One would expect the cost function to be the description length given later in (14). Minimizing (14), however, can result in a difficult optimization problem. This is particularly true for model function approximations where parameters appear nonlinearly. It is more convenient, therefore, to carry out the subset selection using only the likelihood in (5) and then calculating the description length later to rank the models and solve the "argmin" problem for model selection. This approach although sub-optimal is actually quite effective in practice in finding parsimonious but well behaved models.

terms of code length. Judd and Mees [21] apply this principle to discriminate between nonlinear radial basis function models of different size and here we extend these ideas to discriminate between Markov process models.

To formulate the model description length consider the fact that a data set has a certain code length necessary for its description, i.e., the cost of representing the data using floating point number representations. Alternatively, one can consider a model describing the data and calculate the cost of representing the prediction errors of the model plus the code length necessary to represent the model, or its parameters, at a certain precision. We call the code length of the model plus data (errors) the description length. The model with the minimum description length is chosen as the 'best' model. This is the minimum description length principle. In our experience, however, it is better to use information criteria as a screening method to identify a good set of plausible models.

The total description length for a realization z is

$$C(z, \lambda) = C(z|\lambda) + C(\lambda) \tag{8}$$

where the data code length can be approximated by [22]

$$C(z \mid \lambda) = (-\log P(z \mid \lambda)) \approx -\log L(\lambda, z) \tag{9}$$

the negative log likelihood of the data.[3] The code length needed to specify the parameters λ is [21]

$$C(\lambda) \approx \sum_{j=1}^{k} \log \frac{\gamma}{\delta_j}. \tag{10}$$

The parameter γ can be interpreted as the exponent in a floating point representation of the parameters relative to the range of the values of the estimated parameters [26]. The vector δ represents the precision with which the model parameters are specified; however, we will shortly see that these can be *optimized-out* in a consistent manner. In contrast, γ is a free parameter with typical values between 1 and 32. There is actually no strict upper bound on γ but $\gamma = 32$ corresponds to model parameters taking values in the range between 10^{-9} and 10^{9} which seems adequate for almost all practical purposes. (See Appendix A in [21] for a more thorough and technical discussion.) In nonlinear time series reconstruction applications under Gaussian noise assumptions a value of $\gamma = 1$ tends to suggest larger models and a value of $\gamma = 32$ smaller models in terms of the number of basis functions. The model class and likelihood we study is non-Gaussian and so in the examples to follow we study the effect on optimal model choice by considering γ in the range $1 \leq \gamma \leq 32$.

We can bound the description length by considering the maximum likelihood parameter values $\hat{\lambda}$ so that (9) satisfies [22]

$$C(z \mid \lambda) \leq C(z \mid \hat{\lambda}) + \frac{1}{2}\delta^T Q \delta \tag{11}$$

where the Hessian $Q = D_{\lambda\lambda} C(z \mid \hat{\boldsymbol{\lambda}})$. As noted above we eliminate the precisions δ by optimization. Rewrite (8) as

$$C(z, \lambda) \leq C(z \mid \hat{\lambda}) + \frac{1}{2}\delta^T Q \delta + k \log \gamma - \sum_{j=1}^{k} \log \delta_j. \tag{12}$$

The right hand side can be minimized with respect to $\boldsymbol{\delta}$ and the optimal precisions $\hat{\delta}$ are the solution to

$$(Q\delta)_j = 1/\delta_j. \tag{13}$$

This equation can be solved numerically using, for example, Newton's Method. A useful initial condition is $\delta = 1/\sqrt{Q_{ii}}$, which itself is sometimes a good approximation to the solution.

The approximate upper bound to the description length of a model with k parameters can then be written as

$$S_k(z) = -\log L(\hat{\lambda}, z) + \left(\frac{1}{2} + \log \gamma\right) k - \sum_{j=1}^{k} \log \hat{\delta}_j. \tag{14}$$

For each model obtained using the selection methods of the previous section we can calculate $S_k(z)$ and rank the various models. The model with the minimum $S_k(z)$ is chosen as the optimal model.

We note that (14) depends on the number of data in the time series in addition to the number of parameters in the model. In the examples we will investigate this dependence on data length for short time series.

It is interesting to note that, asymptotically, the description length criterion can be simplified to the form [22]

$$M_k(z) = -\log L(\hat{\lambda}, z) + \frac{k}{2} \log N + O(k) \tag{15}$$

where N is the number of data. This expression is essentially the same as the Schwarz Information Criterion (SIC) [29] also known as the Bayesian Information Criterion (BIC). A related criterion for model selection is given by Akaike's (AIC), an information criterion which can be written as [28,22]

$$A_k(z) = -2 \log L(\hat{\lambda}, z) + 2k. \tag{16}$$

Eqs. (15) and (16) are presented solely to place the description length in context and it is (14) which is used and investigated in the examples where the length of the time series is quite short.

4. Examples

We will study the feasibility of the above methods using two examples from population biology. The first example describes simple logistic population growth and the system model belongs to the class of nonlinear models we aim to reconstruct. The second example is a predator–prey system where we attempt to reconstruct the dynamics of the prey population using only observations of the prey population. In this case the true system equations used to generate the time series do not belong to the polynomial model class with which we attempt to model the data. The lack of knowledge of the predator population also requires using the idea of embedding [32] as applied to stochastic signals [33,34] in order

[3] The analogy of coding a model plus errors to picture description length is lost in our example since we do not have a predictive model as such, i.e., a model which provides a pointwise prediction of the time series. However, since the mathematics depends on a likelihood function we can still use the MDL principle to discriminate between competing stochastic process rate models.

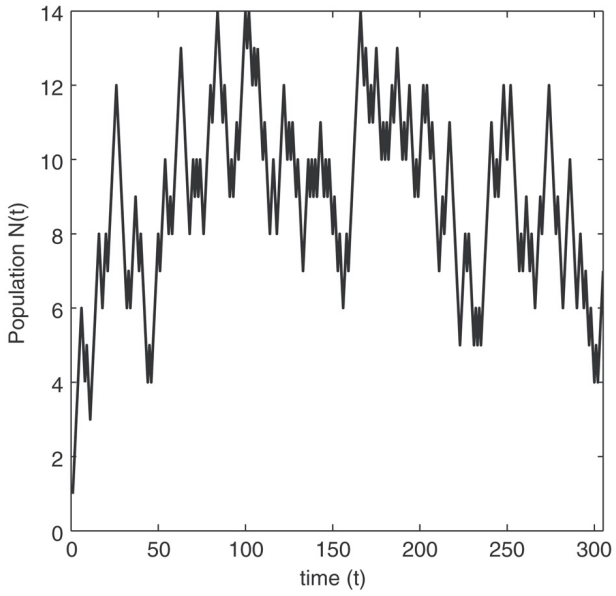

Fig. 1. Typical simulation of the logistic birth–death population model.

to reconstruct a suitable phase space from the prey population measurements.

4.1. Logistic birth–death population model

The logistic model describes general population growth in the absence of immigration and emigration, i.e., only birth and death processes are considered. The birth and death rates are given by polynomial functions of the population, namely

$$B[N(t)] = N(t)(a_1 - b_1 N(t))$$
$$D[N(t)] = N(t)(a_2 + b_2 N(t)).$$

The state space is one dimensional with the state at time t being given by $N(t)$. The population $N(t) \to N(t) + 1$ with probability $B[N(t)]/(B[N(t)] + D[N(t)])$, or, $N(t) \to N(t) - 1$ with probability $D[N(t)]/(B[N(t)] + D[N(t)])$. The time between birth and death events is drawn from an exponential distribution, i.e., $\delta t = -\log(u)/(B[N(t)] + D[N(t)])$ where u is a uniform random variable in (0, 1). A typical simulation [1] with $N(0) = 1$, $a_1 = 2.2$, $a_2 = 0.2$, $b_1 = b_2 = 0.1$ and $t_{\max} = 15$ is shown in Fig. 1.

We assume the process rates can be represented by polynomial functions of maximum order 4 and apply the methods outlined above (cf. the order 2 polynomial in the system model used to generate data belongs to this class of functions). In Table 1 we show the model representation selected as the best model using (14) for different lengths of data sets and values of γ. An entry of Table 1 is interpreted as follows: (x^A, y^B) represents a birth rate model with x basis terms selected with A the highest order of polynomial basis function. y^B describes the selected death rate model. This is not a unique representation for each model but does serve to summarize the degree of nonlinearity of the selected model. For example, the representations $a0 + a1 * x + a3 * x^3$ and $a0 + a2 * x^2 + a3 * x^3$ would both be summarized by 3^3. The true model representation takes the form $(2^2, 2^2)$.

Table 1
Size of selected birth–death polynomial models by description length for varying γ and data lengths N

	$N = 50$	$N = 100$	$N = 200$	$N = 300$
$\gamma = 2^0$	$(5^4, 4^3)$	$(4^4, 2^3)$	$(5^4, 3^3)$	$(5^4, 3^3)$
$\gamma = 2^1$	$(1^0, 1^0)$	$(4^4, 2^3)$	$(5^4, 3^3)$	$(5^4, 3^3)$
$\gamma = 2^2$	$(1^0, 1^0)$	$(1^0, 1^0)$	$(4^4, 2^3)$	$(4^4, 2^3)$
$\gamma = 2^3$	$(1^0, 1^0)$	$(1^0, 1^0)$	$(4^4, 2^3)$	$(4^4, 2^3)$
$\gamma = 2^4$	$(1^0, 1^0)$	$(1^0, 1^0)$	$(1^0, 1^0)$	$(4^4, 2^3)$
$\gamma = 2^5$	$(1^0, 1^0)$	$(1^0, 1^0)$	$(1^0, 1^0)$	$(4^4, 2^3)$

(x^A, y^B) is interpreted as follows: in the birth model there were x basis functions selected with A being the highest order of the polynomial basis function. Similarly y^B represents the corresponding values for the selected death model. The true model has the representation $(2^2, 2^2)$.

Examining Table 1 we see that the true model is not chosen as the best model given the data. We see also that for smaller data sets and increasing γ, simpler model representations are favoured by the minimum description length principle. This is reminiscent of the phenomena observed in nonlinear time series reconstruction where larger γ is known to be a stronger penalty for increasing model size.

It is interesting to discuss reasons why the true model is not selected as the best according to MDL despite belonging to the model class under consideration. We are dealing with finite time series and so the approximations which lead to the derivation of (14) are being stretched and may not be entirely valid when we apply them to short data sets.

A second cause of incorrect model identification can be noise corruption in the time series. In this system there is no artificial noise added to the data but noise is inherent in the model itself and the time series are essentially stochastic signals. It is possible to approximate stochastic process models with systems of deterministic ode's using moment closure techniques [5]. One could then consider the data as signals from a deterministic system contaminated with a high level of dynamic noise which would exacerbate selection problems with MDL.

It is well known in problems of reconstructing predictive deterministic models from short noisy time series that description length curves are non-smooth near the minimum and the true model is often not discovered. For example, in Judd [35] it was demonstrated that the best MDL-selected model was often a degenerate model in the sense that it could be rewritten as iterates of the true model, i.e., the fitted model could be expressed as the true model applied to itself a number of times. Despite having more parameters these degenerate models were fitting the particular realization of the system better under a one-step prediction error than the true model due to using more information from the past resulting in a lower description length.

In Fig. 2 we show the values of description length calculated for different models using $\gamma = 1$ and sets of data of differing lengths. The value of (14) for the true model representation is shown by the dotted line in each case. We see that the true model representation is not the model with minimum description length but does give a MDL value close to the minimum and so we suggest using MDL as a screening

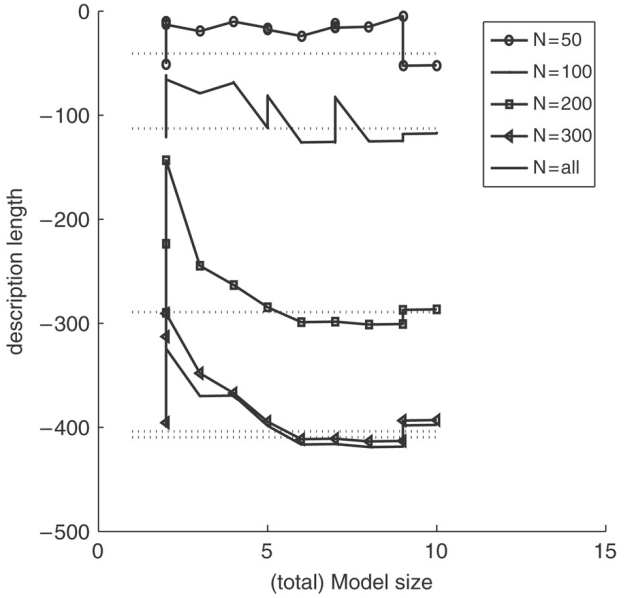

Fig. 2. The figure shows the results of using (14) with $\gamma = 1$ for different lengths of data sets presented to the selection algorithm. The horizontal dotted lines indicate the value of description length returned by the true model structure for each data set. We see that in all cases the true model is not the one with minimum description length.

process for candidate models rather than selecting a definite model representation. Furthermore, if the intended application permits, then an approach involving model averaging may be worthwhile.

4.2. Predator–prey population models

In ecological and biological population modelling there is great interest in the dynamics of interacting populations. In particular, much study has been pursued considering predator–prey behaviour of two interacting species. Early representations led to the development of what are now referred to as Lotka–Volterra models [1]. An example of a stochastic Lotka–Volterra-type model is

$$B_1 = X_1(1.5 - 0.01X_1) \quad \text{and} \quad D_1 = 0.0833X_1X_2$$
$$B_2 = 0.01X_1X_2 \quad \text{and} \quad D_2 = 0.25X_2$$

where B_1 and D_1 represent the birth and death rates of the prey population X_1, and B_2 and D_2 respectively represent the birth and death rates of the predator population X_2. This particular representation and set of parameters gives rise to stochastic cycles of the populations which can be sustained for several periods [1]. A typical realization of the prey population is shown in Fig. 3.

Of interest in ecology is trying to discover representations for the birth and death processes of one species in the absence of knowledge of the competing species. The ideas on process rate reconstruction presented in this paper may provide a partial step forward in addressing this question. We attempt to reconstruct the behaviour of the prey population from an observation of a realization of the above stochastic Lotka–Volterra model. The observations will consist only of time series for the prey

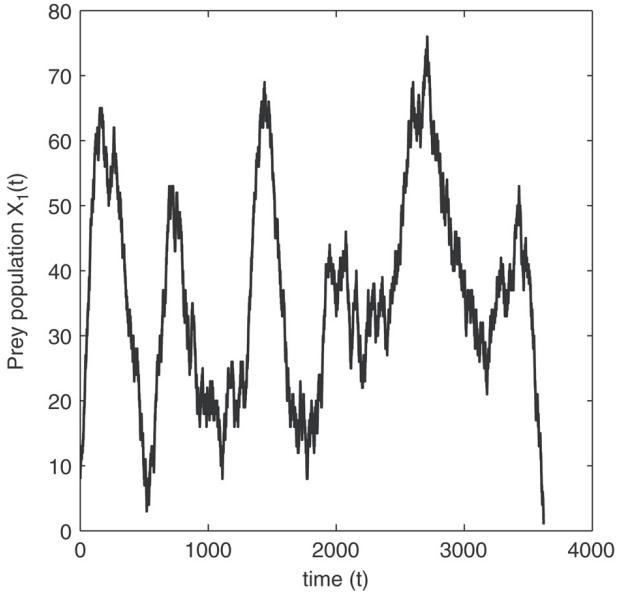

Fig. 3. A typical simulation of the prey population of a Lotka–Volterra-type model.

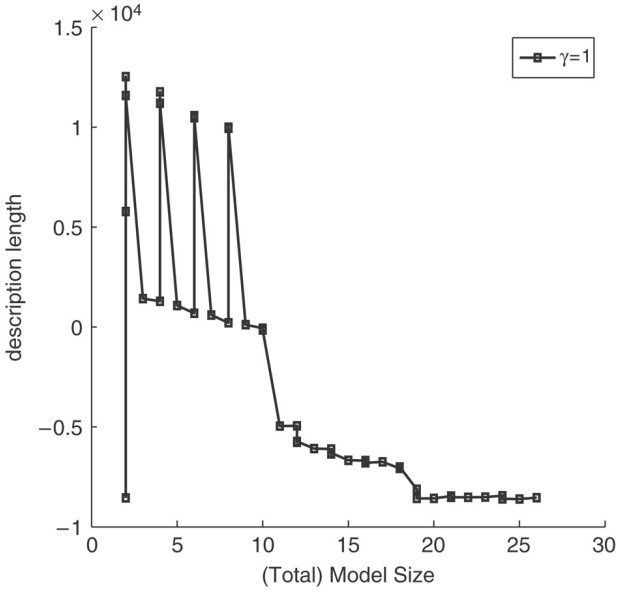

Fig. 4. The model selection criterion (14) with respect to model size with $\gamma = 1$. The MDL model is a nonlinear representation with a total of 25 basis functions.

population. We will consider polynomial representations of the prey birth and death process rates and consider $s(n) = (X_1(n - 150), X_1(n))$ as the system state. The lag of 150 is chosen to be approximately 1/4 the number of data points in a typical cycle. In this embedded space, for example, possible 3^2 representations include $X_1(n) + X_1(n - 150) + X_1(n)^2$ and $a_0 + X_1(n - 150) + X_1(n - 150)^2$.

In Fig. 4 we show the results of considering polynomial representations of the process rates with maximum order 4 and calculating (14) with $\gamma = 1$. The model selected using MDL has total model size equal to 25 and is represented by $(14^4, 11^4)$, a nonlinear representation with degree 4

polynomials in both the reconstructed prey population birth and death process rates.

5. Conclusion

Markov processes are widely used in modelling population dynamics, epidemiology, chemical reactions and systems of interacting agents. Here we have demonstrated how process rates within this stochastic framework can be approximated using polynomial functions. The problem of model selection was addressed by adapting methods based on information criteria used in nonlinear time series reconstruction. The methods were illustrated using two idealized examples from population biology. We developed methods of model selection for approximating the process rates of Markov process models focusing on the case of complete data. The calculation of the likelihood is much harder in the case of missing data and new methods must be developed. Nonetheless, it is anticipated that much of the framework for model selection presented here will be applicable when only incomplete observations are available. The scope for modelling systems using a stochastic framework is wide and we believe the methods introduced here can aid their diverse application.

Acknowledgments

This work was funded by the Scottish Executive Environment and Rural Affairs Department (SEERAD).

References

[1] E. Renshaw, Modelling Biological Populations in Space and Time, Cambridge, 1991.
[2] B. Bolker, S.W. Pacala, Using moment equations to understand stochastically driven spatial pattern formation in ecological systems, Theor. Popul. Biol. 52 (1997) 179–197.
[3] J.H. Matis, T.R. Kiffe, On the cumulants of population size for the stochastic power law logistic model, Theor. Popul. Biol. 53 (1998) 16–29.
[4] D.A. Rand, Correlation equations and pair approximations for spatial ecologies, in: J. McGlade (Ed.), Advanced Ecological Theory: Principles and Applications, Blackwell Science, Oxford, 1999, p. 100.
[5] M.J. Keeling, Metapopulation moments: coupling, stochasticity and persistence, J. Anim. Ecol. 69 (2000) 725–736.
[6] M.J. Keeling, Multiplicative moments and measures of persistence in ecology, J. Theor. Biol. 205 (2000) 269–281.
[7] V. Isham, Assessing the variability of stochastic epidemics, Math. Biosci. 107 (1991) 209–224.
[8] N.G. Van Kampen, Stochastic Processes in Physics and Chemistry, North Holland, Amsterdam, 1992.
[9] D.R. Cox, H.D. Miller, The Theory of Stochastic Processes, Chapman and Hall, London, 1965.
[10] G. Marion, D.L. Swain, M.R. Hutchings, Understanding foraging behaviour in spatially heterogeneous environments, J. Theor. Biol. 232 (2005) 127–142.
[11] P. Whittle, On the use of the normal approximation in the treatment of stochastic processes, J. Roy. Statist. Soc. B 19 (1957) 268–281.
[12] P. Holmes, E. Brown, J. Moehlis, R. Boyacz, J. Gao, P. Hu, G. Aston-Jones, E. Clayton, J. Rajkowski, J.D. Cohen, Optimal decisions: From neural spikes, through stochastic differential equations, to behavior, in: Proceedings 2004 International Symposium on Nonlinear Theory and its Applications, NOLTA2004, vol. 1, 2004, pp. 15–18.
[13] C.J.F. ter Braak, R.S.B. Etienne, Improved Bayesian analysis of metapopulation data with an application to a tree frog metapopulation, Ecology 84 (2003) 231–241.
[14] P.D. O'Neill, G.O. Roberts, Bayesian inference for partially observed stochastic epidemics, J. Roy. Statist. Soc. A 162 (1999) 121–129.
[15] G.J. Gibson, Investigating mechanisms of spatio-temporal epidemic spread using stochastic models, Phytopathology 87 (1997) 138–146.
[16] G.J. Gibson, Markov chain Monte Carlo methods for fitting and testing spatio-temporal stochastic models in plant epidemiology, Appl. Stat. 46 (1997) 215–233.
[17] G. Streftaris, G.J. Gibson, Bayesian analysis of experimental epidemics of foot-and-mouth disease, Proc. Roy. Soc. B 271 (2004) 1111–1117.
[18] G.J. Gibson, E. Renshaw, Likelihood estimation for stochastic compartmental models using Markov chain methods, Statist. Comput. 11 (2001) 335–346.
[19] J.G. Propp, D.B. Wilson, Exact sampling with coupled Markov chains and application to statistical mechanics, Random Structures Algorithms 9 (1996) 223–252.
[20] M.K. Cowles, B.P. Carlin, Markov chain Monte Carlo convergence diagnostics: A comparative review, J. Amer. Statist. Soc. 91 (1996) 883–904.
[21] K. Judd, A. Mees, On selecting models for nonlinear time series, Physica D 82 (1995) 426–444.
[22] J. Rissanen, Stochastic Complexity in Statistical Inquiry, World Scientific, 1989.
[23] K. Judd, A. Mees, Embedding as a modeling problem, Physica D 120 (3–4) (1998) 273–286.
[24] M. Small, C.K. Tse, Minimum description length neural networks for time series prediction, Phys. Rev. E 66 (2002) 066701–1–066701–11.
[25] M. Small, K. Judd, Detecting periodicity in experimental data using linear modeling techniques, Phys. Rev. E 59 (1998) 1379–1385.
[26] T. Nakamura, Modelling nonlinear time series using selection methods and information criteria, Ph.D. Thesis, University of Western Australia, 2003.
[27] T. Nakamura, K. Judd, A.I. Mees, M. Small, A comparative study of information criteria for model selection. Internat. J. Bifur. Chaos 16(7) (2006) (in press).
[28] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Control 19 (6) (1974) 716–723.
[29] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[30] D.M. Walker, G. Marion, Approximating stochastic process rates with nonlinear functions, in: Proceedings 2004 International Symposium on Nonlinear Theory and its Applications, NOLTA2004, vol. 1, 2004, pp. 191–194.
[31] M. Parlar, Interactive Operations Research with Maple, Birkhauser, 2000.
[32] F. Takens, Detecting strange attractors in turbulence, in: D. Rand, L.S. Young (Eds.), Dynamical Systems and Turbulence, Warwick 1980, Springer, 1981, p. 366.
[33] J. Stark, D.S. Broomhead, M.E. Davies, J. Huke, Takens embedding theorems for forced and stochastic systems, Nonlinear Anal. 30 (1997) 5303–5314.
[34] J. Stark, D.S. Broomhead, M.E. Davies, J. Huke, Delay embeddings of forced systems: Ii stochastic forcing, J. Nonlinear Sci. 13 (2003) 519–577.
[35] K. Judd, Building optimal models of time series, in: Gouesboet et al. (Eds.), Chaos and its Reconstruction, Nova, 2003.