

VC-dimension and Structural Risk Minimization for the analysis of nonlinear ecological models

Giorgio Corani^a Marino Gatto^a

^a*Politecnico di Milano, Dipartimento di Elettronica e Informazione
corani(gatto)@elet.polimi.it*

PREPRINT TO APPEAR ON:
Applied Mathematics and Computation, 2006

Abstract

The problem of distinguishing density-independent (DI) from density-dependent (DD) demographic time series is important for understanding the mechanisms that regulate populations of animals and plants. We address this problem in a novel way by means of Statistical Learning Theory. First, we estimate the VC-dimensions of the best known nonlinear ecological models through the methodology proposed by Vapnik et al. (1994). Then, we generate noisy artificial time series, both DI and DD, and use Structural Risk Minimization (SRM) to recognize the model underlying the data from among a suite of alternative candidates. The method shows an encouraging ability in distinguishing between DI and DD time series.

Key words: VC-dimension estimation, Structural Risk Minimization, model selection, demographic time series, density dependence

1 Introduction

A widely addressed problem in ecology is the identification of the basic mechanisms underlying the observed course of population abundances. In its simplest form, the problem is to identify a suitable relationship of the type $N_{t+1} = f(N_t)$, where N_t is the total abundance at time t . T. R. Malthus, the founder of modern demography, in his famous work of 1798 [1] proposed a simple linear model:

$$N_{t+1} = \lambda N_t \tag{1}$$

which yields the geometric growth $N_t = \lambda^t N_0$. Depending on whether λ (*intrinsic finite rate of increase*) is larger or smaller than 1, the population increases indefinitely or tends to extinction. The main assumption underlying the Malthusian model is that the environment can provide each individual with the same amount of resources necessary to survival and reproduction, regardless of the population density. This is actually called the *density-independence* hypothesis.

However, no population grows indefinitely; as the density rises, some competition takes place between individuals (for example for food, water, or reproduction), slowing down or halting the population increase. In this case, density-independence is no longer a suitable assumption; the Malthusian model becomes inadequate and nonlinear models must be introduced. The two simplest nonlinear demographic models are those proposed by three great fish biologists: Beverton and Holt on one side and Ricker on the other. The Beverton-Holt model [2] assumes a hyperbolic decrease of the growth rate with density and is given by

$$N_{t+1} = \frac{\lambda N_t}{(1 + \alpha N_t)} \quad \alpha > 0 \tag{2}$$

For $\lambda > 1$ it is characterized by a positive stable equilibrium which is approached without oscillations.

In 1948 the Canadian biologist William E. Ricker introduced a simple density-dependent demographic model [3,4] which assumes an exponential decrease of the demographic rate. It is given by

$$N_{t+1} = \lambda N_t \exp(-bN_t) \tag{3}$$

The model was lucidly investigated by Beverton and Holt themselves [2] who remarked that, depending on λ , the population can reach a steady state with

or without oscillations, undergo permanent oscillations of constant period and amplitude or undergo what they termed irregular and violent oscillations. The nature of this dynamic complexity was not fully understood by ecologists until Robert May [5,6] introduced the basic concepts of bifurcation and chaos into population ecology. In fact, the Ricker model is characterized by a Feigenbaum cascade [7] for increasing $a = \ln \lambda$. It is worthwhile to remark that both the Beverton Holt and the Ricker model contain the Malthusian model as a special case. Parameters α and b in the two models account for the effects of competition between individuals belonging to the same population.

The Beverton-Holt model can be generalized [8] by introducing a power coefficient γ as:

$$N_{t+1} = \frac{\lambda N_t}{(1 + \alpha N_t)^\gamma} \quad (\text{generalized Beverton I}) \quad (4)$$

or

$$N_{t+1} = \frac{\lambda N_t}{(1 + \alpha N_t^\gamma)} \quad (\text{generalized Beverton II}) \quad (5)$$

For $\gamma > 1$ these two models have a dynamic behavior similar to the Ricker one.

All the populations are influenced by exogenous forcing which is best described by a stochastic process in most cases. The simplest way to include environmental fluctuations in the models is through an additive, uncorrelated noise, namely

$$N_{t+1} = f(N_t) + n_t \quad (6)$$

where n_t is a random variable with a suitable probability distribution. Biologically, n_t can represent random errors in measuring N_{t+1} or random migration (in or out the population boundaries).

To recognize whether a population is growing in a density-dependent or independent way is of great practical importance in the design of proper policies for sustainable management and exploitation of populations. In particular, the problem of making the best prediction of N_{t+1} from the observed abundances N_t is of paramount interest. Therefore, a great research effort has been devoted over the past two decades to statistically distinguish density-dependent from independent time series. A milestone in this context is the work of Dennis and Taper [9], who proposed a powerful hypothesis-testing framework for model identification based on parametric bootstrapping of likelihood ratios. However, a general weakness of hypothesis testing frameworks is that they contrast a

single density-independent model (usually the Malthusian one) with a single alternative density-dependent one (usually the Ricker model), while it would be desirable to investigate the suitability of several alternative models.

To overcome such a limitation, several authors used information criteria, which indeed allow one to choose the best from a suite of several alternative models. In particular, the Schwartz Information Criterion (SIC) is most used by ecologists [10–13], since it was shown to detect the model structures underlying the data better than the Akaike Information Criterion (AIC) [14]. However, information criteria have major drawbacks too, because of the very restricting hypotheses assumed for their formulation. First, they can be used just with linear candidate models, or with nonlinear models that can be somehow linearized (e.g., taking the log-densities in the Ricker model). Second, they are based on asymptotic arguments and, as such, they are really suited to work properly with large datasets, while in population ecology short time series are usually available.

A possible way to tackle these problems is to use Statistical Learning Theory. In fact, with reference to regression problems, it has been shown [15] that the VC-based model selection framework, called Structural Risk Minimization (SRM), can consistently overperform traditional Information Criteria for different dataset sizes and noise levels; remarkably, strong advantages are observed on smaller datasets and higher noise levels, which are usual conditions in ecological modelling.

Based on this results, in this work we address the density-dependence detection problem via the SRM approach. In particular, we estimate the VC-dimension of nonlinear ecological models, in order to subsequently exploit the SRM as model selection framework. More precisely, we will compare different models in terms of their ability to predict N_{t+1} from N_t . To this end, we will consider the prediction problem as a regression (linear or nonlinear) of N_{t+1} against N_t .

However, estimating VC-dimensions of nonlinear regressors is a challenging task. Up to now, SRM has been used mainly with linear estimators, because VC-dimensions are almost always unknown for nonlinear models. Here we use a simulation methodology proposed by Vapnik et al. [16] for estimating the VC-dimension of any learning machine; significant algorithmic improvements were then carried out in [17]. These improvements are implemented to estimate the VC-dimension of demographic predictors.

Estimating the VC-dimension of density-dependent models is interesting *per se*, since it provides an index of dynamic complexity of classes of different models; for example, the Ricker model (which can generate periodic oscillations and chaos) is much more flexible than the Beverton-Holt, even though they

have the same number of parameters. Indeed, the Beverton-Holt model will never be able to fit oscillatory or chaotic data because it entails a stable steady state under any parameter choice. The calculated differences of VC-dimensions will provide quantitative evaluations of such complexity differences.

Nevertheless, the main interest of estimating the VC-dimensions is that it allows the application of SRM to the problem of density dependence detection. To test the effectiveness of SRM we perform a huge number of simulations of both density-dependent and independent models, using many different parametric settings and noise levels to corrupt the simulated data; for each simulated dataset, SRM is applied to choose a model from among the five candidates presented above. The performances of SRM in recognizing the real model underlying the data are satisfactory, especially for discriminating between density independent and dependent time series. It should be remarked that the VC-dimensions are estimated once and for all; therefore, this will allow ecologists to use SRM as model selection framework with a minimum effort in the future.

The paper is organized as follows: section 2 describes the VC-dimension estimation methodology, while section 3 presents the VC-dimensions estimates obtained for the nonlinear ecological models. Finally, section 4 presents the results of the density-dependence detection test based on SRM.

2 VC-dimension estimation methodology

The VC-dimension h of a binary classifier is defined as the maximum number of vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_h$ that can be shattered (i.e., separated in all the 2^h possible ways) by the classifier itself. Such a definition does not provide a methodology for measuring it; indeed, VC-dimension is actually known with precision only for linear classifiers and estimators and for a few other models. However, a method to estimate the VC-dimension of any learning machine has been proposed by Vapnik et al. [16]. We provide here just an outline of the methodology; theoretical proofs and more detailed explanations can be found in the original work.

The methodology was originally defined with reference to binary classifiers; nevertheless it can be easily extended to regressors as well [18], using a few attentions. In binary classification, each d -dimensional input sample $\mathbf{x} = (x_1, x_2, \dots, x_d)$ needs to be labeled as belonging either to class zero (i.e., $y = 0$) or to class one (i.e., $y = 1$). The accuracy of a classifier for a dataset of size n is measured by the misclassification error rate ϵ , defined as:

$$\epsilon = \frac{n_{\text{misclassified}}}{n} \tag{7}$$

The VC-dimension estimation methodology analyzes the maximum deviations $\xi(n)$ of error rates between two independently labeled datasets $Z^1 = \{(\mathbf{x}_1^1, y_1^1), \dots, (\mathbf{x}_n^1, y_n^1)\}$ and $Z^2 = \{(\mathbf{x}_1^2, y_1^2), \dots, (\mathbf{x}_n^2, y_n^2)\}$ of size n :

$$\xi(n) = \max_{\omega} \left(\left| \epsilon(Z^1) - \epsilon(Z^2) \right| \right) \quad (8)$$

where ω is the set of parameters of the binary classifier.

According to the theoretical findings of Vapnik et al. [18], $\xi(n)$ is bounded as follows:

$$E[\xi(n)] \leq \Phi(n/h) \quad (9)$$

where Φ is a function defined as:

$$\Phi\left(\frac{n}{h}\right) = \begin{cases} 1 & \text{if } \left(\frac{n}{h} < .5\right) \\ a \frac{\ln(2\frac{n}{h})+1}{\frac{n}{h}-k} \left(\sqrt{1 + \frac{b(\frac{n}{h}-k)}{\ln(2\frac{n}{h})+1}} + 1 \right) & \text{otherwise} \end{cases} \quad (10)$$

Parameters a, b, k , are thought to be constant and universal; they are set to 0.16, 1.2, 0.14298 [16].

Since bound (9) is tight, one can safely assume

$$\Phi\left(\frac{n}{h}\right) \approx E[\xi(n)] \quad (11)$$

By means of formulas (10) and (11) we can obtain an estimate of h fitting the empirical average of the maximum error rate deviation $\xi(n)$ to $\Phi\left(\frac{n}{h}\right)$.

To measure $\xi(n)$, the error rate has to be minimized on dataset Z^1 , and maximized on Z^2 . This can be accomplished through the following procedure:

- (1) Generate a set of size $2n$, $Z = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{2n}, y_{2n})\}$ and split it into two sets Z^1 and Z^2 of size n . Both \mathbf{x} and y are generated randomly: the \mathbf{x} 's as a Gaussian noise, and the y 's by Bernoulli trials with $p = \frac{1}{2}$.
- (2) Flip the labels of set Z^2 , transforming the j -th input-output pair as:

$$\begin{cases} \mathbf{x}_j : \text{unchanged} \\ \bar{y}_j = (1 - y_j) \end{cases} \quad (12)$$

thus obtaining the set $\overline{Z^2}$.

- (3) Consider the set $\overline{Z} = \{Z^1, \overline{Z^2}\}$. Train the classifier on the whole dataset \overline{Z} , i.e. find ω^* that minimizes $\epsilon(\overline{Z})$.
- (4) Measure the maximum error rate deviation $\xi(n)$ between Z^1 and Z^2 by using ω^* to calculate the error rates $\epsilon(Z^1)$ and $\epsilon(Z^2)$:

$$\xi(n) = \epsilon(Z^1) - \epsilon(Z^2) \quad (13)$$

The procedure summarized in steps 1-4 is defined as an *experiment*; different experiments are repeated for each of m different sample sizes (*design points*) n_1, n_2, \dots, n_m . Usually, 20 different experiments are performed at each design point. Estimates $\xi_{avg}(n_i)$ of $E[\xi(n_i)]$ are obtained by averaging the $\xi(n_i)$ resulting from experiments.

The VC-dimension estimate h^* is finally determined by fitting formula (10) of Φ to the m average deviations rates $\xi_{avg}(n_i)$:

$$h^* = \arg \min_h \sum_{i=1}^m [\xi_{avg}(n_i) - \Phi(n_i/h)]^2 \quad (14)$$

If the model under investigation is a regressor (i.e., $y = g(x, \omega)$, with g being a real valued function), the procedure must be suitably modified. In particular, we can still use the same set \overline{Z} with binary y 's, but the training at step 3 is performed by minimizing the squared deviation

$$\sum_{i=1}^n (y_i - g(\mathbf{x}_i, \omega))^2 + \sum_{i=n+1}^{2n} (\overline{y}_i - g(\mathbf{x}_i, \omega))^2 \quad (15)$$

Moreover, to calculate the error rates at step 4, the real valued estimates $\hat{y} = g(\mathbf{x}, \omega^*)$ are converted to binary classes by applying the Heaviside step function $H(0.5 - \hat{y})$. Thus, it is possible to compute regression errors as misclassification errors and to evaluate the ϵ error rates and the corresponding deviation $\xi(n)$.

A problem that arises with nonlinear regression is the choice of suitable design points. Vapnik et al. introduced design points as ratios $(\frac{n}{h})$, within the range $0.5 \leq (\frac{n}{h}) \leq 30$. However, such an approach is feasible only if h is known a priori; indeed, results reported in [16] refer just to linear estimators. Considering nonlinear models, ratios (n/h) are a priori unknown; therefore, we adopt an iterative approach, using a first guess h^0 (equal to the number of parameters of each model) to define the m design points as $\{(n_1 = 0.5 * h^0), \dots, (n_m = 30 * h^0)\}$. A first estimate h^1 is then obtained by means of formula (14); it is used to define a new set of design points $\{(n_1 = 0.5 * h^1), \dots, (n_m = 30 * h^1)\}$, in correspondence of which a new estimate h^2 is returned. The procedure lasts until the obtained estimate h^{i+1} matches the value h^i used for building the design points.

The procedure to estimate the VC-dimension of nonlinear regressors is simply applied to the demographic models by taking N_{t+1} and N_t as the dependent and the independent variable respectively. VC-theory is based on \mathbf{x}_i and y_i being independent identically distributed random variables, which does not hold for time series. [19] has outlined an approach that extends Vapnik’s method of Structural Risk Minimization to time series generated by an underlying mixing stochastic process. However, this approach requires the knowledge of the mixing rate of the process, which is not at all easy to estimate. Therefore, we will straightforwardly use the standard SRM for nonlinear regressors and our heuristic approach will be justified only a posteriori by the results for the problem of model choice. If density independence can be correctly distinguished from density dependence with a high probability, we can claim that our approach is valid.

The procedure described above for nonlinear regressors returns a real-valued estimate of the VC-dimension. This estimate can then be rounded to the closest integer, if we refer to the theoretical definition of VC-dimension for the shattering problem, or retained as is, if we consider it just as an index of model complexity for analytical model selection (Cherkassky, personal communication).

2.1 Experiment distribution optimization

Since there is no theoretical justification to perform an equal number of experiments at each design point, i.e., to use a uniform experiment distribution, some algorithmic modifications have been proposed in [17], which return a non uniform distribution of experiments between the design points. In this case the fitting error is defined as:

$$MSE = E [(\xi(n) - \Phi(n, h))^2] \quad (16)$$

to account for non uniform experiment distribution.

The algorithm for designing the optimized experiment distribution is initialized using a first-guess experiment distribution (for example a uniform one) and then proceeds as follows:

- (1) Use the current experiment distribution and compute the fitting error (16);
- (2) Rank all the design points n_1, n_2, \dots, n_d according to the quantity:

$$C(n_i) = \frac{MSE(\text{remove } n_i) - MSE}{k(n_i)} \quad (17)$$

where $MSE(\text{remove } n_i)$ is the fitting error given all the design points other than n_i . The number of experiments at n_i , denoted as $k(n_i)$, is used as normalization factor in order to return the specific MSE contribution of each experiment at the given design point. The quantity $C(n_i)$ is hence an estimate of the contribution of design point n_i to the overall MSE; it is likely to be negative when the variance of $\xi(n)$ is large at n_i , and vice versa.

- (3) Change the experiment distribution by removing an experiment from the worst design point (i.e., having lowest C) and adding it to the best design point (i.e., having highest C).
- (4) Estimate h^* and evaluate again the fitting error (16): if the change leads to a bigger MSE, it is reversed and the experiment is then added to the 2nd (3rd, 4th etc.) best design point, until one point returns a smaller MSE;
- (5) Repeat step 1-4 until one of the two termination criteria is met:
 - (a) all the design points give non-positive contribution
 - (b) all the design points that have a positive contribution are *saturated*. A design point is defined as saturated if it collects 25% of the number of total experiments; no further experiments can be added to saturated points, in order to keep a wide enough set of design points.

2.2 Results on low-order linear regressors

Estimation results reported in both [16] and [17] consider linear regressors having at least VC-dimension 10; however, the presented ecological models are likely to have much smaller VC-dimensions. Therefore, we test the above described methods with low-order linear regressors, in order to assess the estimation accuracy for models of limited complexity.

A linear regressor of order k , formalized as $y = \sum_{i=1}^k w_k x_k + w_0$, is known to have VC-dimension $k + 1$. According to the results reported in Table 1, the optimized procedure always returns the correct estimate, while the uniform design is often characterized by an error of one unit. Therefore, we decide to adopt the optimized design procedure also for estimating the VC-dimensions of the ecological models.

3 VC-dimension estimation for ecological models

According to Cherkassky (personal communication) it is in principle correct to use real-valued VC estimates (i.e., as they are returned from the estimation algorithm) for analytic model selection; thus, in the sequel we analyze the

h	h^*	h^*
(true VC)	(uniform design)	(optimized design)
2	2	2
3	3	3
4	3	4
5	4	5
6	5	6

Table 1

True VC dimensions (h) and estimates (h^*) obtained for linear regressors.

obtained real-valued estimates.

Since real-valued estimates may slightly vary between different estimation trials, we adopt a two-step approach to improve the estimation:

- (1) use of the optimized design procedure, which returns an optimal distribution of experiments and a corresponding VC-dimension estimate h^* ;
- (2) computation of 29 further estimates of h^* by using the optimal experiment distribution identified at step 1.

The average of the 30 different estimates of h^* is retained as final VC-dimension estimate.

The parameters estimation of a nonlinear model requires the use of iterative algorithms (in particular, we use the Nelder Mead algorithm implemented in the Matlab TM function `fminsearch`) which minimize the training error starting from a randomly provided parameters initialization. To avoid being trapped in local minima, the identification algorithm is started several times, using different initializations. Increasing the number of initializations allows the improvement -up to a certain extent- of the model fit to training data. Thus, the generalization ability of a given nonlinear model is partly determined by the number of initializations used for estimating the model parameters. With regard to the VC estimation methodology, increasing the number of initializations is likely to decrease the misclassification rates on the training data $\epsilon(Z^1)$ and to increase that on the testing data $\epsilon(Z^2)$; this would lead to greater error deviations ξ . We expect therefore that increasing the number of initializations would result in an increase of the VC estimates. In fact, this is what occurs as demonstrated by the results presented in Figures 1a-1d. Estimates seem to stabilize using about 40 initializations; models with three parameters, which require a much more difficult search of the minimum for the estimation algorithms, take greater advantage from repeated initializations.

Summing up, we have these final estimates (40 initializations) for the VC

$\frac{n}{h}$	<i>Bev.</i>	<i>Ricker</i>	<i>Bev. I</i>	<i>Bev. II</i>	$\frac{n}{h}$	<i>Bev.</i>	<i>Ricker</i>	<i>Bev. I</i>	<i>Bev. II</i>
.5	15	19	17	17	3.8	5	5	5	5
.65	5	6	5	5	5	21	18	20	20
.8	5	2	5	5	6.5	20	21	20	19
1	5	4	5	5	8	21	22	20	21
1.2	5	4	5	6	10	25	25	25	25
1.5	5	4	5	5	15	61	61	61	60
2	4	6	5	4	20	81	81	82	81
2.8	6	5	5	5	30	81	82	80	82

Table 2

Optimized experiment distribution of different ecological models. They tend to favor design points corresponding to larger datasets, as already reported for linear estimators in [17].

dimensions of ecological models: 1.36 for Beverton-Holt, 1.79 for Ricker, 3.00 for generalized Beverton-Holt I, and 3.32 for generalized Beverton-Holt II. The coefficients of variations of the 30 estimates of the different models are in the range 0.06 – 0.1. The optimal experiment distributions obtained for ecological models are given in Table 2.

It is worth noticing that the model with the simplest dynamics (Beverton-Holt) has the lowest VC-dimension, which is actually closer to 1 than to the number of free parameters (two). Generalized Beverton-Holt models are significantly more complex than Ricker. Actually, the increase of VC-dimensions is larger than the increase of the number of free parameters (from two to three). Generalized Beverton-Holt II has a dimension larger than the number of free parameters.

4 SRM model selection

The estimation of VC-dimensions for regressors can be used to select the appropriate model within the framework provided by Statistical Learning Theory.

Denoting as h the VC-dimension of a class of functions g , Structural Risk Minimization (SRM) provides an analytic upper bound on the prediction risk

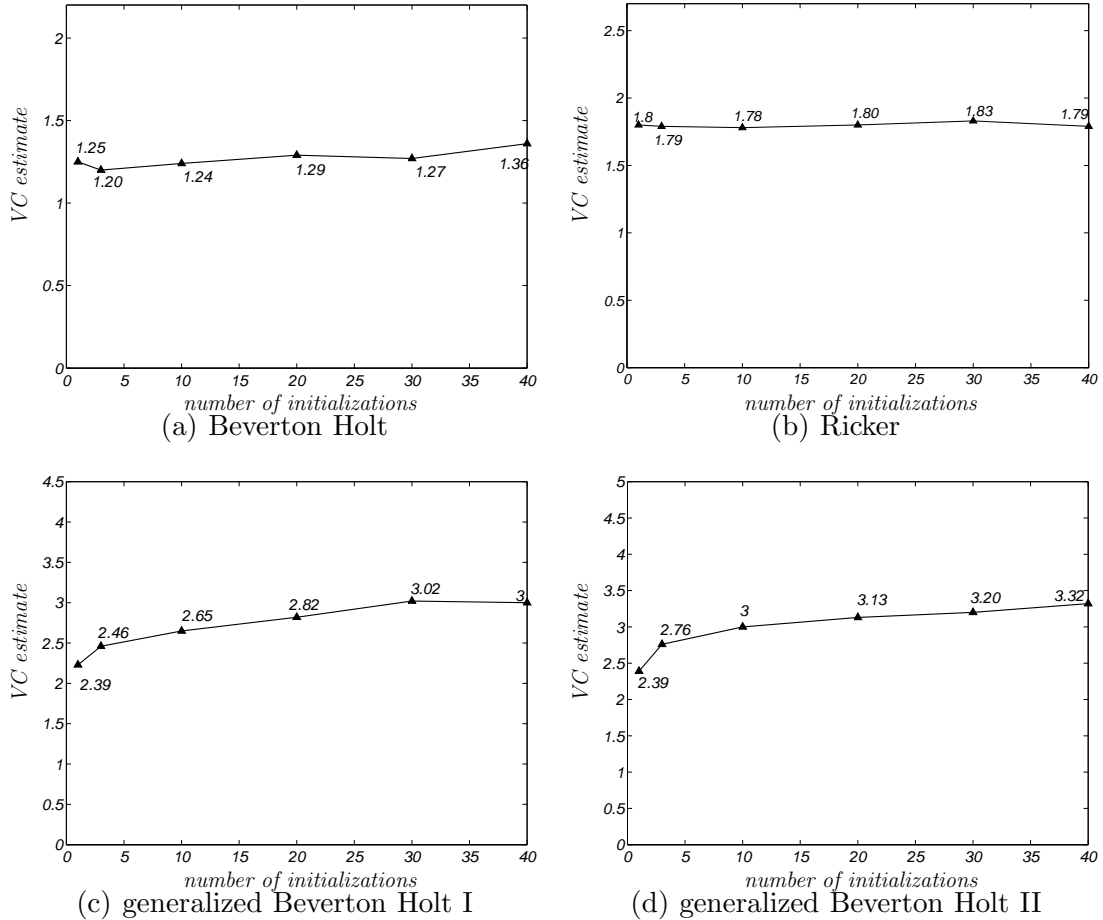


Fig. 1. VC-dimension estimates show an increasing trend with the number of parameters initializations during training. This is most apparent with generalized Beverton-Holt models, which display a difference of up to 40% between the estimates obtained using 1 and 40 initializations.

which can be written [15] as:

$$R(\omega) \leq \frac{1}{n} \sum_1^n (y_i - g(\mathbf{x}_i, \omega))^2 \left[1 - \sqrt{p - p \ln p + \frac{\ln n}{2n}} \right]_+^{-1} \quad (18)$$

where $p = h/n$. Such a form of the bound is obtained from the general analytical bound derived within the VC-theory, by properly setting the values of the theoretical constants for “practical” regression problems [15]. As such, the VC-bound is derived under very general assumptions (finite sample, non-linear settings). A model is finally chosen among a set of candidates when it minimizes the upper bound for the prediction risk. We now proceed to testing the power of the SRM criterion for selecting the appropriate ecological model, using the VC dimensions estimated previously.

	<i>Malth.</i>	<i>Bev.</i>	<i>Ricker</i>	<i>Bev. I</i>	<i>Bev. II</i>
<i>average performance</i>	78%	22%	0%	0%	0%
<i>sensitivity to λ</i>					
$\lambda = 1.00$	57%	43%	0%	0%	0%
$\lambda = 1.05$	79%	21%	0%	0%	0%
$\lambda = 1.11$	86%	14%	0%	0%	0%
$\lambda = 1.28$	91%	9%	0%	0%	0%
<i>sensitivity to δ</i>					
$\delta = 0.05$	81%	19%	0%	0%	0%
$\delta = 0.10$	77%	23%	0%	0%	0%
$\delta = 0.25$	76%	23%	0%	0%	0%
<i>sensitivity to q</i>					
$q = 16$	70%	30%	0%	0%	0%
$q = 64$	87%	13%	0%	0%	0%

Table 3

Results for the detection of density-independent Malthusian demography. The values in each cell represent the percentage a given model is chosen by means of SRM.

4.1 Density-independent time series

In this first series of experiments, we simulate the Malthusian model, corrupting the simulation with noise. The simulation setting which characterizes a Malthusian simulation is given by the following parameters:

- the finite rate of increase λ ;
- the noise level δ ;
- the simulation length q .

To corrupt the simulation with additive noise, we adopt the following approach:

- (1) a deterministic simulation of q steps is performed, using equation $N_{t+1} = \lambda N_t$;
- (2) the noisy simulation of q steps is performed using equation $N_{t+1} = \lambda N_t + \delta \bar{N} W N$, where $W N$ denotes a standard white noise ($\mu = 0, \sigma = 1$), and \bar{N} the average population abundance of the deterministic simulation. Thus, the noise level δ is defined as the ratio of the noise standard deviation to \bar{N} . Negative abundances, which are possible because of the additive noise, are set to 0 in order to preserve the biological meaning of N_t .

We use 24 different simulation settings, given by all the possible combinations of $\lambda = [1; 1.05; 1.28]$, $q = [16; 64]$, $\delta = [0.05; 0.1; 0.25]$; for each setting, 500 different simulations are performed, starting from the initial condition $N_0 = 20$. In order to be coherent with the VC estimates experimental settings, identification algorithms of nonlinear models are initialized 40 times.

Table 3 shows a summary of the results; it does not report the details regarding all the 24 different simulation settings, rather it provides the average model choice percentages for wide classes of settings.

In the global average, the SRM test correctly recognizes the Malthusian model about 78% of the times; the remaining 22% of the times it chooses Beverton-Holt, which is the nonlinear model with the lowest VC-dimension. Higher percentages of correct model choice are obtained with higher λ (population increasing at a faster rate) and longer time series. The test success is rather robust to the increase of noise level.

4.2 *Density-dependent time series*

In a second experiment, we simulate the generalized Beverton-Holt I model. The simulation settings are in this case defined by:

- the set of model parameters λ, γ, α . We use four parameter sets, each corresponding to one of the four different dynamic behaviors of the model: stable equilibrium without oscillations, damped oscillations, limit cycle and chaos [8].

In analogy with the previous case, simulation settings comprise also:

- the noise level δ , whose values are set to $[0.05; 0.1; 0.25]$;
- the simulation length q , whose values are set to $[16; 64]$.

Simulations have been corrupted with additive noise, as done in the Malthusian case.

Table 4 summarizes the results. First of all, we notice that the density dependence detection goal is satisfactorily met, since the Malthusian model is chosen just 2% of the times. On the other hand, the recognition of the correct density-dependent model is a bit disappointing, since the generalized Beverton-Holt model I is chosen just 12% of the times. While the percentage a model is chosen does not show a great sensitivity to noise or to time series length, it does depend on the underlying dynamic behavior of the model. In particular, if the simulated data are generated by parameters that correspond to equilibrium with monotonic damping, the simple Beverton-Holt model ($h^* = 1.36$)

	<i>Malth.</i>	<i>Bev.</i>	<i>Ricker</i>	<i>Bev. I</i>	<i>Bev. II</i>
<i>average performance</i>	2%	25%	50%	12%	11%
<i>sensitivity to model dynamics</i>					
monotonic damping	6%	88%	0%	6%	0%
damped oscillations	0%	13%	49%	18%	20%
limit cycle	0%	1%	67%	16%	16%
chaos	0%	0%	85%	7%	8%
<i>sensitivity to δ</i>					
$\delta = 0.05$	0%	21%	49%	15%	15%
$\delta = 0.10$	1%	27%	54%	9%	9%
$\delta = 0.25$	4%	29%	47%	11%	9%
<i>sensitivity to q</i>					
$q = 16$	2%	25%	51%	13%	9%
$q = 64$	1%	26%	50%	11%	13%

Table 4

Results of the detection test for the density-dependent demography (generalized Beverton-Holt I model).

is chosen most of the times. Also, this is the only situation where the Malthusian model is chosen -though very rarely-. With more complex dynamics, the Malthusian model is never chosen, and the percentage the simple Beverton-Holt model is chosen becomes much lower. The generalized Beverton-Holt models are chosen with almost identical percentages, ranging between 10% and 20% depending on the simulation settings. The Ricker model is the most widely chosen, and remarkably the percentage increases as the simulated dynamics become more complex.

5 Conclusions

We have estimated, for the first time, the VC-dimensions of nonlinear ecological models and used them to tackle the problem of distinguishing density-independent demography from density-dependent.

The obtained results are undoubtedly encouraging, because the correct recognition percentages are about 78% and 98%, respectively.

Thus, one can use SRM with the VC-dimensions estimates provided here if the problem of density-dependence detection in demographic time series is to be

faced. One should be aware, though, that there is a non negligible probability of mistaking Malthusian population growth for density-dependent growth.

The experiments described here are not directly comparable with any previous work in the ecological literature, since no previous density detection frameworks manages a suite of both linear and nonlinear models. However, we have good reasons to be confident that SRM can be a much more powerful approach than previous density-dependence detection methodologies. In a series of further experiments, we have compared the performances of SRM and parametric bootstrapping [9] to distinguish between Malthusian and Ricker models. For density-independence detection, parametric bootstrapping and SRM show very close performances, but SRM greatly overperforms parametric bootstrapping in density dependence recognition. The advantage is strongly emphasized in critical situations, such as smaller datasets, higher noise levels, initial condition set for simulation close to the Ricker model equilibrium.

A further development of our work will be the inclusion of multiplicative noise in the demographic model. This kind of noise is more apt to describing random variations of demographic parameters such as survival or fertility. In this case, it is natural to consider the regression of the logarithm of the finite growth rate against the population abundance. Of course, the VC-dimensions for these regressors are different, even if they correspond to the same ecological models. The results of applying SRM to this case are therefore nontrivial.

References

- [1] T. Malthus, *An Essay on the Principle of Population*. 1798.
- [2] R. Beverton and S. Holt, *On the Dynamics of Exploited Fish Populations*. H.M. Stationery Office, 1957.
- [3] W. Ricker, “Stock and recruitment.,” *J. of the Fisheries Research Board of Canada*, vol. 11, pp. 559–623, 1954.
- [4] W. Ricker and R. Foerster, “Computation of fish production,” *Bull. Bingham Oceanogra. Coll*, vol. 11, pp. 173–211, 1948.
- [5] R. May, “Biological populations with nonoverlapping generations: stable points, stable cycles and chaos.,” *Science*, vol. 186, pp. 645–647, 1974.
- [6] R. May and G. Oster, “Bifurcations and dynamic complexity in simple ecological models,” *American Naturalist*, vol. 110, pp. 573–599, 1976.
- [7] Y. Kuznetsov, *Elements of Applied Bifurcation Theory*. Springer Verlag, 1998.
- [8] M. Begon, J. Harper, and C. R. Townsend, *Ecology : Individuals, Populations, and Communities*. Blackwell Science In, 1998.

- [9] B. Dennis and M. Taper, “Density dependence in time series observation of natural populations: estimation and testing,” *Ecological Monographs*, vol. 64, no. 2, pp. 205–244, 1994.
- [10] D. Strong, A. Whippler, A. Child, and B. Dennis, “Model selection for a subterranean trophic cascade: root-feeding caterpillars and entomopathogenic nematodes,” *Ecology*, vol. 80, no. 8, pp. 2750–2761, 1999.
- [11] B. Dennis and M. Otten, “Joint effects of density dependence and rainfall on abundance on san joaquin kit fox,” *J. Wild. Management*, vol. 64, no. 2, pp. 388–400, 2000.
- [12] Z. Zeng, R. Nowiersky, M. Taper, B. Dennis, and W. Kemp, “Complex population dynamics in the real world: Modeling the influence of time-varying parameters and time lags,” *Ecology*, vol. 79, no. 6, pp. 2193–2209, 1998.
- [13] M. Taper and P. Gogan, “The northern yellowstone elk: density dependence and climatic conditions,” *J. Wild. Management*, vol. 66, no. 1, pp. 106–122, 2002.
- [14] M. Hooten, *Distinguishing forms of statistical density dependence and independence in animal time series data using Information Criterion*. PhD thesis, Montana State University, 1995.
- [15] V. Cherkassky, X. Shao, F. Mulier, and V. Vapnik, “Model complexity control for regression using VC generalization bounds,” *IEEE Trans. on Neural Networks*, vol. 10, no. 5, pp. 1075–1089, 1999.
- [16] V. Vapnik, E. Levin, and Y. Cun, “Measuring the VC-dimension of a learning machine,” *Neural Computation*, vol. 6, pp. 851–876, 1994.
- [17] X. Shao and V. Cherkassky, “Measuring the VC-dimension using optimized experimental design,” *Neural Computation*, vol. 12, no. 8, pp. 1969–1986, 2000.
- [18] V. Cherkassky and F. Mulier, *Learning from data*. Wiley-Interscience, 1998.
- [19] R. Meir, “Nonparametric time series prediction through adaptive model selection,” *Machine Learning*, vol. 39, pp. 5–34, 2000.