AN APPLICATION OF STRUCTURAL RISK MINIMIZATION TO THE SELECTION OF ECOLOGICAL MODELS

Giorgio Corani * **Marino Gatto** *

* *Dipartimento di Elettronica ed Informazione*
Politecnico di Milano
e-mail:corani@elet.polimi.it

Abstract: The problem of distinguishing density-independent (DI) from density-dependent (DD) demographic time series has been addressed in the past via hypothesis testing based on parametric bootstrapping (PBLR) and, in later works, by Information Criteria such as FPE or SIC. Here, we address the problem in a novel way using Structural Risk Minimization (SRM). DI and DD time series corrupted with noise are extensively simulated using a drift (DI) and a Ricker (DD) model; on each generated time series, both models are identified, and then one is selected by FPE, SIC and SRM. The probability of density-[in]dependence recognition is statistically assessed and compared with the results obtained via PBLR in a previous work.

Keywords: Time series analysis, Identification, Ecology, Learning Control

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 proposed a simple linear model $N_{t+1} = \lambda N_t$, which yields the geometric growth $N_t = \lambda^t N_0$. By taking logarithms, the equation becomes

$$\ln\left(\frac{N_{t+1}}{N_t}\right) = a \qquad (1)$$

where $a = \ln(\lambda)$. This model is usually referred to in demography as *drift model.* Depending on a, the population increases indefinitely or tends to the 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 density rises, 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 and the drift model becomes inadequate. One of the most flexible models describing density dependence is the one introduced in 1948 by the Canadian biologist William E. Ricker:

$$ln\left(\frac{N_{t+1}}{N_t}\right) = a + bN_t \;\; (a > 0, b < 0) \qquad (2)$$

It is worth noting that the only nontrivial equilibrium of such a model corresponds to $\overline{N} = -a/b$. This equilibrium is stable for $a < 2$, and there are damped oscillations for $1 < a < 2$.

Since the recognition of density-dependence is of great practical importance in the design of proper policies for sustainable management and exploitation of natural populations, this topic stimulated a great research effort over the past three decades. Earlier works were based on hypothesis-testing approaches, the two hypotheses being usually constituted by the drift and the Ricker model. For instance, Dennis and Taper proposed a powerful hypothesis testing framework based on parametric bootstrapping of likelihood ratios (PBLR) (Dennis and Taper, 1994). However, as hypothesis testing suffered sometimes from the problem of low power, several authors (Taper and Gogan, 2002) proposed the use of information criteria (IC) to choose the best among a suite of alternative models including both density-independent and density-dependent demographies. The Final Prediction Error (FPE) and in particular the Schwartz Information Criterion (SIC) appear to be the most widely used by ecologists. It is worthwhile to mention that traditional IC's are based on asymptotic arguments, which therefore hold just for large datasets, and assume a set of common hypotheseses such as the linearity of both the target function underlying the data and the approximating functions used as models, which do not hold in real case studies.

As a viable alternative to classical IC, we propose the use of the model selection criterion developed within Statistical Learning Theory (SLT) called Structural Risk Minimization (SRM) (Vapnik, 1995). SLT is a modelling framework of great generality, which works with finite samples without assuming any particular condition about the data, or the class of the approximating functions. The very core of SLT is the concept of VC-dimension h, a complexity index for classes of functions; for the comprehension of the proposed application it is enough to know that in the linear case, VC-dimension corresponds to the number of free parameters of the model. On the contrary, VC-dimensions of nonlinear models are generally unknown, and this constitutes in fact a major obstacle to a wide application of SLT findings.

With reference to linear regression problems, it has been shown (Cherkassky *et al.*, 1999) that SRM can consistently overperform traditional Information Criteria (SIC, FPE, etc.) for different dataset sizes and noise levels, with stronger advantages for smaller datasets and higher noise levels, which are quite usual conditions in ecological modelling.

In our experimental framework, we generate noisy artificial time series by using both the drift and the Ricker model. For each generated time series, both models are identified and one of them is selected according to FPE, SIC or SRM. We have designed our simulation experiments consistently to (Dennis and Taper, 1994), because our aim is to compare our results with PBLR, which is very well assessed in ecological modelling.

The paper is organized as follows: Section 2 details the model selection approaches, Section 3 describes the experimental methodology, Section 4 and 5 illustrate the results obtained for the recognition of the density-independent and density-dependent demography, Section 6 describes an application to 3 populations of large mammals.

2. THE MODEL SELECTION PROBLEM

From an abstract viewpoint we can think of the model selection problem as the problem of approximating the functioning of a true system; such a system receives an input vector $\mathbf{x}$, characterized by a probability distribution $P(\mathbf{x})$ and correspondingly returns an output y, according to the conditional distribution $P(y|\mathbf{x})$. Both $P(\mathbf{x})$ and $P(y|\mathbf{x})$ are *unknown.* We assume that the system is represented by the unknown relationship:

$$y = g(\mathbf{x}) + \epsilon \tag{3}$$

where ϵ is an independent identically distributed zero mean random noise.

A model selection procedure is aimed at choosing the best approximating function among a set of several candidates $f_j(\mathbf{x}, \omega)$, where ω denotes the parameters specifying the function, and the subscript j refers to one of different classes of functions. For example, class j might be a polynomial of degree j.

The choice is based on a finite number q of samples $(\mathbf{x_i}, y_i), i = 1, \ldots, q$. If, as usual, the quality of the approximation is measured through the squared error, the optimal approximating function should in principle minimize the following *prediction risk functional*:

$$R_j(\omega) = \int (y - f_j(\mathbf{x}, \omega))^2 dP(\mathbf{x}, y) \tag{4}$$

which is however unknown because the joint probability distribution function $P(\mathbf{x}, y) = P(y|x)P(x)$ is unknown.

On the other hand, what can be experimentally measured by using the q samples is the *empirical risk*:

$$R_j(\omega)_{emp} = \frac{1}{q} \sum_{i=1}^{i=q} (y_i - f_j(\mathbf{x}_i, \omega))^2 \tag{5}$$

Information Criteria attempt to estimate the unknown prediction risk (4) as the known empirical risk (5), penalized by some measure of the model complexity. Once an accurate estimate of the prediction risk is found, the model that minimizes the estimated prediction risk with respect to both the class j of functions and the parameters defining each function inside the class is chosen. In general, for a function f_j having d_j free parameters, ICs take the form:

$$\text{estimated risk}(f_j) = R_j(\omega)_{emp}\ r\left(p_j\right) \tag{6}$$

where $r(p)$ is the penalization function and p_j denotes the ratio d_j/q. In this paper we consider the following Information Criteria:

$$\text{FPE estimated risk}(f_j) = R_j(\omega)_{emp}\left[\frac{(1+p_j)}{(1-p_j)}\right] \tag{7}$$

$$\text{SIC estimated risk}(f_j) =$$
$$= R_j(\omega)_{emp}[1+\frac{\ln(q)}{2}p_j(1-p_j)^{-1}] \tag{8}$$

These classical approaches are motivated by asymptotic arguments ($q \to \infty$) for linear models and indeed risk estimates provided by FPE and SIC are asymptotically equivalent. They also assume that the target function $g(\mathbf{x})$ is contained in the set of candidate approximating functions $f_j(\mathbf{x},\omega)$. It is worthwhile to note that the experiments performed in this paper with artificial time series will actually satisfy such an assumption, which is not met in real world case studies.

As for the PBLR approach, it can briefly summarized as follows: it is a hypothesis test, where model i is contrasted against model j. The test statistic is the ratio Λ_{ij} of the likelihood function L_i maximized over the parameters values of model i, to the likelihood L_j, also maximized over the parameters of model j. The decision is made in favor of Model i if $\Lambda_{ij} > c$, where c is a cutoff value selected so that the probability of wrongly choosing Model i when data arise from Model j is fixed at a small number, known as *test size.* Under the PBLR approach, the cutoff value is estimated via parametric bootstrapping (Dennis and Taper, 1994).

These classical approaches can be contrasted with the VC-theory approach where, for a sample of finite length q, one can calculate a bound for the risk functional (4). For "practical" regression problems, the following inequality holds with probability $\left(1-\frac{1}{\sqrt{q}}\right)$(Cherkassky *et al.*, 1999):

$$R_j(\omega)) \leq$$

$$\leq R_j(\omega)_{emp}\left[1-\sqrt{p_j - p_j\ln p_j + \frac{\ln(q)}{2q}}\right]_+^{-1} \tag{9}$$

where $p_j = \frac{h_j}{q}$ (h_j is the VC-dimension of the j-th class of functions). If the models are linear, h_j coincides with the number of free parameters, so $p_j = \frac{d_j}{q}$. The SRM approach consists in choosing the model that minimizes the right-hand-side of (9). Therefore, in practice SRM is yet another way of penalizing the empirical risk R_{emp}.

With reference to our application, the problem of predicting the rate of demographic increase between year t and year $t+1$ can be obtained by setting

$$\begin{cases} y = ln(\frac{N_{t+1}}{N_t}) \\ x = N_t \end{cases}$$

3. THE MODEL SELECTION EXPERIMENTAL METHODOLOGY

To test the model selection criteria, we generate artificial noisy time series, adopting a log-normal noise:

$$N_{t+1} = N_t\exp(a + bN_t + nZ_t) \tag{10}$$

where n is a parameter defining the noise level and Z a standard normal white noise ($\mu = 0, \sigma^2 = 1$). Coefficient b is clearly set to 0 when the drift model is simulated.

An ensemble of stochastic simulations is characterized by the following set of parameters, which constitute the *simulation setting*:

- the initial condition N_0;
- the model coefficients (a, b);
- the noise level n;
- the simulation length q.

The simulation settings investigated for each model have been designed consistently to (Dennis and Taper, 1994), in order to allow for a coherent comparison of the results of the various criteria with PBLR. The experimental model selection methodology, repeated 500 times for each simulation setting, is as follows:

(1) *stochastic simulation*: perform a q-steps noisy simulation by means of equation (10), using the current simulation setting;
(2) *identification*: estimate the parameters of the Ricker and the drift model by means of linear least squares;
(3) *acceptability check:* as for the PBLR methodology, discard the Ricker model if the estimate of b is positive, and in this case automatically select the drift model for all the

criteria. In fact, b should be a negative parameter, since intraspecific competition negatively affects the population growth rate;

(4) *model selection*: choose the best model according to FPE, SIC and SRM.

4. DENSITY-INDEPENDENCE DETECTION

n	*FPE*	*SIC*	*SRM*
0.05	88%	100%	98%
0.55	72%	99%	94%
1.1	66%	99%	92%
1.6	63%	99%	92%
a	*FPE*	*SIC*	*SRM*
0.05	57%	98%	91%
0.55	70%	99%	94%
1.1	79%	99%	96%
1.6	83%	99%	96%
q	*FPE*	*SIC*	*SRM*
10	71%	96%	90%
20	72%	100%	93%
40	73%	100%	97%
60	74%	100%	98%
average	*FPE*	*SIC*	*SRM*
	72%	99%	94%

Table 1. Percentages of correct detection of the drift demography for different levels of model parameters. Percentages are averaged over all the simulations that share the parameter value shown in the first column.

Simulation settings adopted for the drift model have been obtained by combining in all the possible ways the following values:

- $a = [0.05; 0.55; 1.1; 1.6]$;
- $n = [0.05; 0.55; 1.1; 1.6]$;
- $q = [10; 20; 40; 60]$;
- $N_0 = [64]$.

We use therefore 64 different simulation settings, for a total of 32000 simulations.

Within the PBLR hypothesis testing framework, the drift and the Ricker model constitute respectively the null and the alternative hypothesis. The test size, i.e. the probability of rejecting the drift model when it really underlies the data, has been set to 5% by the authors (Dennis and Taper, 1994). Their extensive simulation of the drift model demonstrated that the effective size of the PBLR test actually met its nominal size.

By contrast, Information Criteria and SRM do not deal with test size, or similar concepts; simply, one chooses the model with the lower estimated risk.

The outcomes of model detection for FPE, SIC, SRM are given in Table 1. The correct recognition percentages of SIC and SRM are close to 100% and are in practice insensitive to any variation in the simulation settings; SIC is slightly more advantageous than SRM. Instead, FPE is sistematically worse (20-30 percentage points lower on average); its effectiveness suffers from high noise levels or low values of a, while it is quite insensitive to the dataset size.

Finally, we remark that about 30% of times the Ricker model is discarded because of a positive estimate of parameter b.

5. DENSITY-DEPENDENCE DETECTION

It is well known (Dennis and Taper, 1994) that parameter b does not influence the probability of recognizing density-dependence as long as it is not zero; its numerical value is a scale factor, which reflects the units in which the population is measured. In fact, setting $P_t = bN_t$ in the Ricker model equation, we obtain $\ln \frac{P_{t+1}}{P_t} = a + P_t$. Therefore, we fixed $b = -0.01$ in all the simulations.

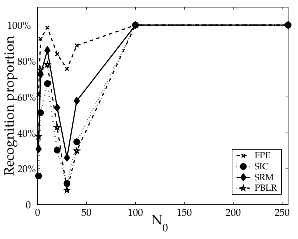


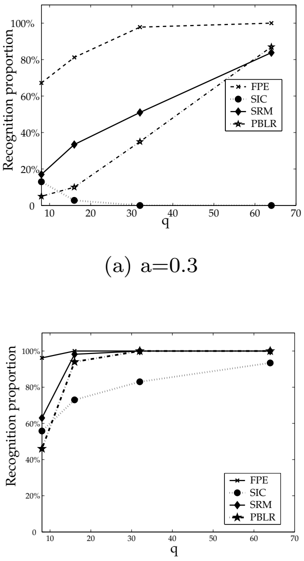
Fig. 1. Proportion of correct detection for the Ricker density-dependent demography as a function of the initial density N_0 (case with $a = 0.3, \overline{N} = 30$).

The first issue investigated is the effect of the initial population size N_0 on the model recognition effectiveness. Coherently with (Dennis and Taper, 1994), we adopted the following simulation settings:

- $a = [0.3;\ 1.2]$;
- $b = [-0.01]$;
- $n = [0.05]$;
- $q = [10]$
- $N_0 = [10; 20; 30; 30; 40; 100; 130; 150; 200; 250]$.

Figure 1 reports the results in the case a=0.3. All the criteria show a similar behavior, from a qualitative point of view: in particular, the percentage of correct detection is minimum when the starting condition N_0 is close to $\overline{N} = -a/b$. In fact, for both values of a, $\overline{N}$ is a stable equilibrium and the density-dependent model moves toward the equilibrium. Therefore, small deviations of the initial population from the equilibrium do not allow the exploration of the dynamical characteristics of the model, thus making the recognition more difficult. When we compare the effectiveness of the model selection approaches, we see that FPE is best, followed by SRM and then by SIC and PBLR. Quite likely, FPE's recognized tendency to overparameterize plays here in a favorable way. When the initial condition is close to the equilibrium - the most difficult situation- it has an advantage of about 40 percentage points over SRM, which has a further advantage of about 25 percentage points over both SIC and PBLR.

On the other hand, correct detection is easier for higher values of parameter a; for $a = 1.2$, for instance, all the criteria behave satisfactorily and also the minimum around $\overline{N}$ is much less apparent. In fact, $a = 1.2$ corresponds to oscillations around $\overline{N}$.



(a) a=0.3

(b) a=1.2

Fig. 2. Ricker model recognition as a function of q and a.

In a further series of experiments, we investigate the effect of the time series length q and the drift parameter a (Figures 2a-b). The simulation settings, coherent with (Dennis and Taper, 1994), are as follows:

- $a = [0.3; 1.2]$;
- $b = [-0.01]$;
- $n = [0.05]$;
- $q = [8; 16; 32; 64]$;
- $N_0 = [-a/b]$.

All the simulations are initialized at the equilibrium, where model recognition is most difficult. As expected, the proportion of correct recognition increases with the time series length with the exception of SIC for small a, whose performance is sursprisingly worse for larger q. FPE performs better than the other criteria specially for very small time-series length. SIC is the worst performing method, while SRM consistently outperforms PBLR, in particular for small q.

In the most critical case, i.e.low a and small dataset, FPE has an advantage of about 40 points over SRM, which additionally outperforms PBLR of 10-20 points; SIC has a recognition percentage close to 0.

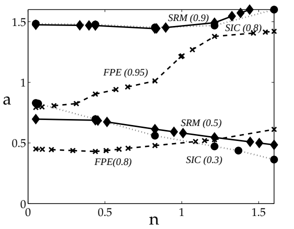


Fig. 3. Contour plots of Ricker model detection proportion as a function of a and q.

In a third series of experiments, we investigate the effect of environmental stochasticity as measured by the noise level n. Such an issue is investigated jointly with finer variations of the drift parameter a. The simulation settings adopted in this case are given by all the possible combinations of the following values:

- $a = [0.05; 0.25; 0.45; 0.8; 1.6]$;
- $b = -0.01$;
- $n = [0.05; 0.25; 0.45; 0.8; 1.6]$;
- $q = 10$;
- $N_0 = -a/b$.

The results are shown in Figure 3 as contour plots of the detection proportion in the parameters plane $n - a$. The most striking feature is that the recognition proportion can increase with the noise level n, with the exception of the FPE criterion. This result may seem counterintuitive. However, such an apparent contradiction can be explained by considering that simulations are started at the model equilibrium; stochastic fluctuations provide deviations from the equilibrium and hence make the correct model selection easier as already noted by Dennis and Taper with reference to PBLR. Why this is not true for FPE is not so easy to explain: quite likely, the FPE tendency to overparameterization (while implying choosing density dependence instead of independence) is somehow hindered by high noise levels.

We can finally conclude that, as for density-dependence recognition, all the criteria share some common features, such as the minimum of power for $N_0 = -a/b$, and the increase of recognition proportion with a and q. The results allow a consistent (i.e., confirmed in all the investigations) conclusion: FPE is in this case the most successful approach, followed by SRM, PBLR, while SIC is certainly worst.

Population	PBLR	FPE	SIC	SRM
G	DI	DI	DI	DI
E1	DD	DD	DI	DD
E2	DD	DD	DD	DD

Table 2. Density-detection case studies.

6. APPLICATION TO FIELD DATA

To compare the various approaches, we apply them to datasets provided in the original paper by (Dennis and Taper, 1994) on PBLR. We try to detect density-dependence in 3 populations of large mammals: the grizzly bear of the Yellowstone region (population G, years 1973-1991), the elk of the Yellowstone region (population E1, years 1968-1979), and the elk of the Grand Teton National Park (population E2, years 1963-1985). First, we identify both the drift and the Ricker model; then we choose one of the two according to the different model selection criteria, and then conclude whether the population is density-dependent or not. The results of these analyses are provided in Table 2. For these populations, the detection of density-[in]dependence is almost consistent among the different criteria, allowing us to conclude that the grizzly population is density-independent, while the two elk populations are density-dependent. Only in the case of population E2, SIC chooses the drift model while all the remaining criteria suggest that the Ricker model is a better choice; according to the results previously presented, it is however reasonable to neglect the SIC indication, which is likely to be too conservative.

The recognition of density-[in]dependence is almost coherent among the different criteria, allowing to conclude that the grizzly population is density-independent, while the two elk populations are density-dependent. Only in the case of population E2, SIC chooses the drift model while all the remaining criteria indicate the Ricker; according to the results previously presented on artificial data, it is however reasonable to neglect such SIC indication, which is likely to be too conservative.

7. CONCLUDING REMARKS

In this work, we address the density-dependence detection problem by comparing the performances provided by the traditional SIC and FPE model selection criteria, the well-established PBLR hypothesis test, and SRM, the model selection criterion developed within the Statistical Learning Theory framework. Although SRM was shown to outperform many traditional model selection criteria (Cherkassky *et al.*, 1999), it has been rarely used in time series analysis up to now.

In our case study, we simulate with noise the simple drift (DI) and Ricker (DD) model under a huge variety of different settings, and then we perform model recognition experiments on the noisy time series, in order to evaluate the ability of the different model selection criteria in distinguishing density-independence from density-dependence.

Our experimental findings show that (i) SIC is "conservatively biased", i.e. it correctly detects density-independence in all but few cases, but the recognition of the more parameterized density-dependent model can be very unsatisfactory. On the other hand, (ii) FPE displays a somewhat opposite characteristic; it is the best performing criterion in recognizing density-dependence, but is the worst in correctly detecting the less parameterized drift model. (iii) SRM and PBLR are better balanced as they provide a very high probability of recognizing density-independence, and at the same time behave satisfactorily also when the Ricker model underlies the time series. However (iv) SRM consistently overperforms PBLR in recognizing density dependence specially with short time series and with simulations started at the model equilibrium. Therefore, despite SRM has been formalized under the assumption of independent identically distributed data, which is not completely true for density and demographic growth rates, we can conclude that it is a promising tool to recognize whether a time series underlies density-dependence or not.

REFERENCES

Cherkassky, V., X. Shao, F. Mulier and V. Vapnik (1999). Model complexity control for regression using VC generalization bounds. *IEEE Trans. on Neural Networks* **10**(5), 1075–1089.

Dennis, B. and M.L. Taper (1994). Density dependence in time series observation of natural populations: estimation and testing. *Ecological Monographs* **64**(2), 205–244.

Taper, M.L. and P.J. Gogan (2002). The northern yellowstone elk: density dependence and climatic conditions. *J. Wild. Management* **66**(1), 106–122.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer-Verlag.