Structural Risk Minimization as model selection criterion for ecological populations

Giorgio Corani

Politecnico di Milano
corani@elet.polimi.it

Acknowledgments:

Marino Gatto (Politecnico di Milano)

Outline

- Demographic models for predicting population abundances
- The model selection problem
- Comparison via simulation of FPE, SIC and SRM
- The *alpine ibex* case study

The problem, from an ecological point of view

- Predicting the future population abundance N_{t+1} from the current observation N_t.

- Usually, one tries to predict the *rate of increase* $Y_{t+1} = \log\left(\frac{N_{t+1}}{N_t}\right)$ instead of N_{t+1}

- Y_{t+1} can depend for instance on:
  - the current population N_t
  - exogenous climatic forcings $X_{1t}, X_{2t}, \ldots X_{mt}$ (rainfall, temperature, etc).

- Reliable population abundances prediction lead to designing proper exploitation policies

Basic demographic models

- The Malthusian ($\mathbf{M}$) model (1798):

$$N_{t+1} = \lambda N_t \Rightarrow y_{t+1} = \ln\left(\frac{N_{t+1}}{N_t}\right) = a \ (a > 0)$$

  The population growths as $N_q = \lambda^q N_0$. The environment is supposed to provide each individual with the same resources (*density-independence*), regardless of the population size.

- The Ricker ($\mathbf{R}$) model (1948) (*density-dependence*):

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

  Depending on the parameters settings, it can reach the stable equilibrium $\overline{N} = -a/b$, or mimick limit cycles or chaos.

Ricker models with covariates

- The Ricker model can be extended including covariates:

$$\ln\left(\frac{N_{t+1}}{N_t}\right) = a + bN_t + cX_{1t} \ (a > 0, b < 0) \ \text{(RI)}$$

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

  Analogously, we obtain models RIII, RIV etc.

- Remark: in practice, we are considering linear regressors

- One usually ends up with a broad suite of demographic models. How to choose among them?

The model selection problem (the machine learning point of view)

- The unknow true system

$$y = g(\mathbf{x}) + \epsilon \tag{1}$$

  is supposed

  - to receive an input vector $\mathbf{x}$, with probability distribution $P(\mathbf{x})$
  - and to return an output y, according to $P(y|\mathbf{x})$.

- Both $P(\mathbf{x})$ and $P(y|\mathbf{x})$ are unknown.

- Notation remark: $P(\mathbf{x}, y) = P(x)P(y|\mathbf{x})$

- A finite number q of observations $(\mathbf{x_i}, y_i), i = 1, \ldots, q$ is available

Model selection problem (II)

- We consider a set of candidate approximating functions $f_j(\mathbf{x}, \theta)$.

- The optimal approximating function should in principle minimize the *risk functional*:

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

which is unknown because $P(\mathbf{x}, y)$ is unknown.

- What can be measured is instead the *empirical risk* (training error):

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

Information Criteria (ICs)

- ICs attempt to estimate the unknown risk functional penalizing the empirical risk. For a function f_j having d_j free parameters, denoting $p_j = d_j/q$, the risk is estimated as:

$$ER_j(\theta) = R_j(\theta)_{emp}\, r\,(p_j)$$

where $r(p)$ is the penalization function.

- FPE (Akaike's Final Prediction Error, 1970):

$$ER_j^{FPE}(\theta) = R_j(\theta)_{emp}\left[\frac{(1+p_j)}{(1-p_j)}\right]$$

- SIC (Schwartz Information Criterion, 1978; aka BIC, aka SC)

$$ER_j^{SIC}(\theta) = R_j(\theta)_{emp}[1 + \frac{\ln(q)}{2} p_j (1-p_j)^{-1}]$$

ICs basic assumptions

- There is a wide set of ICs in literature, obtained under different hypotheses. However, all of them:
  - are motivated by asymptotic arguments ($q \rightarrow \infty$?)
  - assume the linearity of the approximating functions
  - and that $g(\mathbf{x})$ is contained in the set of approximating functions.

- They are therefore often used outside of their constitutive assumptions.

Structural Risk Minimization (Vapnik, 1998)

- SRM is a model selection approach of great generality:
  - the dataset is assumed to be of finite size q;
  - no particular requirements on noise, probability distributions, etc.;
  - no request for linearity.
- For practical regression problems, the following bound holds with probability $\left(1 - \frac{1}{\sqrt{q}}\right)$ (Cherkassky et al., 1999):

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

  where p_j is defined as h_j/q.
- What's h?

VC-dimension

- h_j is the VC-dimension of the approximating function f_j.
- It constitutes an index of complexity for a given function; see (Vapnik, 1998) for a rigorous definition.
- In the linear case, it corresponds to the number of free parameters.
- In the non linear case, it can be estimated by the algorithms proposed in Vapnik (1994) and Cherkassky (2000).
- See (Corani and Gatto, 2005) for an attempt to estimate the VC-dimension of the commonest nonlinear ecological models.
- In the talk, we will deal just with linear models.

The idea of the comparison between ICs and SRM

- The idea:
  - simulate stocastichally (i.e., with noise) different demographic models, under a wide variety of parametric settings;
  - on each noisy simulation, identify a broad set of models;
  - choose one candidate via FPE, SIC, SRM;
  - evaluate the generalization of the chosen models on testing set;
  - repeat the procedure to collect a statistically significant dataset

- Remark: we must set:

$$\begin{cases} y = ln(\frac{N_{t+1}}{N_t}) \\ \mathbf{x} = [N_t, X_{1t}, \ldots, X_{kt}] \end{cases}$$

Stochastic simulation details

- Models are stochastically simulated as

$$N_{t+1} = N_t \exp(a + bN_t + cX_{1t} + dX_{2t} + nZ_t)$$

where Z_t is a WN[0,1].

- Depending on the simulated models, b, c, d can be set to 0.

- For each different *simulation setting*, 500 different simulations are performed

- 96 (2*3*4*4) simulation settings for model R:
  - $N_0 = [100; -a/b]$, $a = [.5; 1; 1.5]$, $b = -0.01$, $n = [.05; .1; .25; .5]$, $q = [10; 20; 50; 100]$;

- A similar variety of simulation settings is used for the remaining models.

The picture of the comparison methodology

1. perform 500 $q-$steps *simulations* using the current setting;

2. *identification* of candidates (M, R, RI, RII, **RIII, RIV**) via linear LS;

3. discard density-dependent models with $\hat{b} < 0$;

4. *model selection* according to FPE, SIC and SRM;

5. *generalization assessment*:

   (a) compute stochastichally 20 times N_{q+1} and then y_{q+1}
   (b) use the chosen models to deterministically compute $\hat{y}_{q+1}$
   (c) compute the error statistics for each criterion ($20 \bullet 500 = 10,000$ samples)

Results: model choices

Aggregated percentages of correct recognition:

FPE				SIC			
M	R	RI	RII	M	R	RI	RII
75%	**60%**	**51%**	**52%**	**71%**	**55%**	**49%**	**51%**

SRM			
M	R	RI	RII
98%	**89%**	**69%**	**67%**

- SRM correctly selects each model with the highest frequency
- Both SIC and FPE generally tend, when failing, to overparameterized models
- On the contrary SRM tends, when failing, to too simple models.

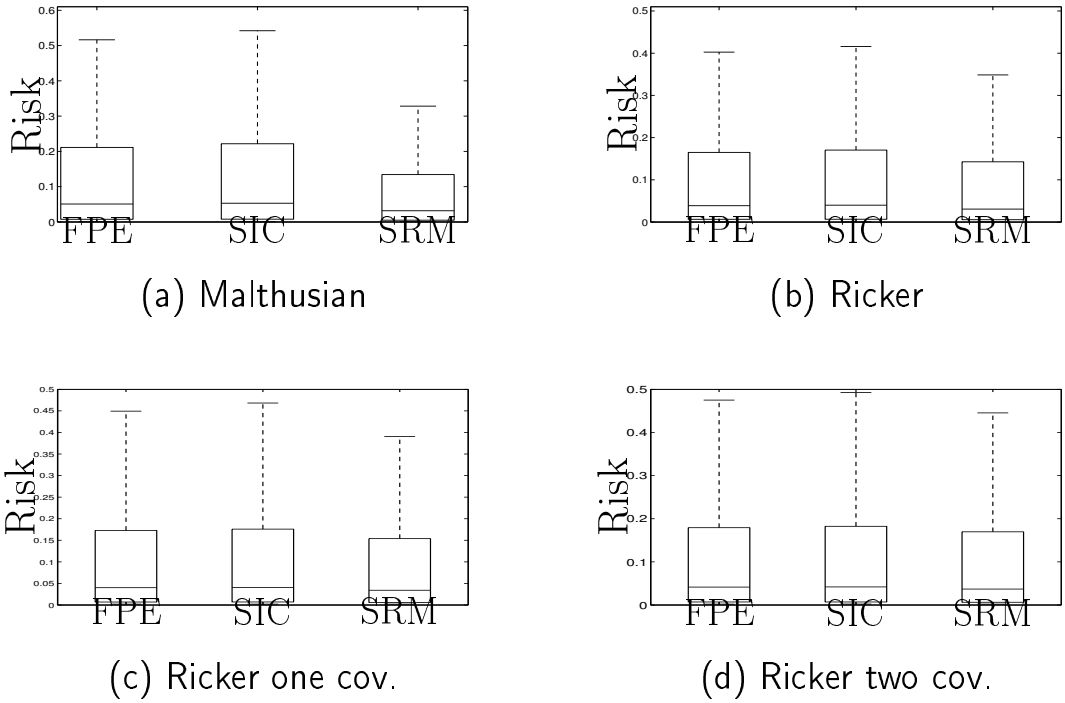
Sensitivity to the dataset size q

	Malthusian model		
	FPE	SIC	SRM
q			
10	62%	50%	**95%**
20	76%	68%	**96%**
50	81%	81%	99%
100	82%	86%	100%

	Ricker model		
	FPE	SIC	SRM
q			
10	24%	34%	**65%**
20	48%	50%	**85%**
50	68%	52%	97%
100	78%	54%	99%

- On shorts datasets, SRM clearly outperforms the traditional asymptotical criteria
- Such a finding is confirmed for **all** the simulated models

Out-of-sample prediction error: risk analysis



(a) Malthusian (b) Ricker

(c) Ricker one cov. (d) Ricker two cov.

Figure 1: Box and whiskers plots of prediction risk

- SRM achieves in every case the lowest risk on each model
- The gap will increase on multi-step predictions

Alpine ibex case study

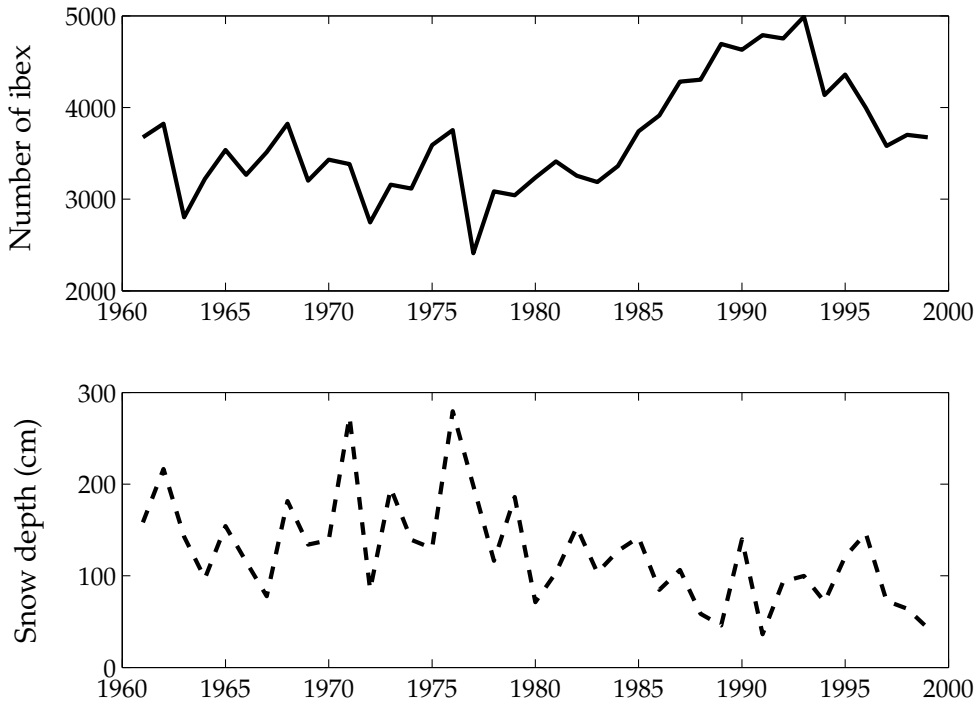
- Population of the Gran Paradiso National Park (Italy)

- Hunting not allowed, large predators absent. Population dynamics can be explained via density-dependence and climate forcings

The dataset

- Censuses and meteorology over 1960-2000 (Jacobson et al., Ecology, 2004)

- Snow depth as the most significant climate driver (Jacobson et al., Ecology, 2004)



Main findings of Jacobson et al. (Ecology, 2004)

- They consider as possible regression variables N_t, $L_t = \ln(N_t)$, the snow depth S_t, the products $S_t L_t$ and $S_t N_t$.

- By cross-checking different statistical tests, they select the threshold model:

$$y = a + \begin{cases} c_1 S_t + d_1 N_t S_t & \text{if } S_t < \overline{S} \\ c_2 S_t + d_2 N_t S_t & \text{if } S_t > \overline{S} \end{cases}$$

where $\overline{S} = \mu(S_t) + .5\sigma(S_t) = 154\,\text{cm}$

Analysis by SRM

- Identification of all the models considered in the previous paper, both linear (not reported here) and with threshold.

	Threshold models (unique a)							
model #	const	N_t	L_t	S_t	N_tS_t	L_tS_t	DoF	SRM
15	*	*		*	*		**7**	**0.0120**
16	*		*	*		*	**7**	**0.0121**
17	*			*	*		**5**	**0.0140**
18	*			*		*	**5**	**0.0140**
19	*	*			*		5	0.0270
20	*		*			*	5	0.0284
21	*	*		*			5	0.0357
22	*		*	*			5	0.0304
23	*				*		3	0.0258

- model 17 was selected in Jacobson et al. (Ecology, 2004)

LOO-CV with the oustanding candidates

Model #	$(y_t - \hat{y}_t)^2$	$\sqrt{(N_t - \hat{N}_t)^2}$
15	.0046	43
16	.0047	43
17	.0062	46
18	.0064	47

- remark: $\hat{N}_{t+1} = N_t \exp(\hat{y}_{t+1})$
- LOO-CV confirms the ranking of SRM, leading to choose model 15.

A novel subset of models

- There is no particular reason for a having a unique estimate. We then introduce models where a is estimated twice.

- General performances improvement (six models out of 9 improve their SRM score)

model #	const	N_t	L_t	S_t	N_tS_t	L_tS_t	DoF	SRM
24	*	*		*	*		8	0.0129
25	*		*	*		*	8	0.0129
⇒26	*****			*****	*****		**6**	**0.0107**
⇒27	*****			*****		*****	**6**	**0.0108**
⇒28	*****	*****			*****		**6**	**0.0110**
⇒29	*		*			*	6	0.0114
⇒30	*	*		*			6	0.0114
⇒31	*		*	*			6	0.0114
32	*					*	4	0.0236

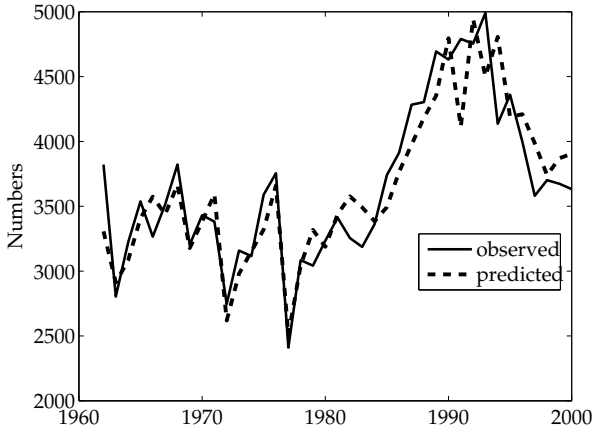
LOO-CV with the ultimate candidates

Model #	$(y_t - \hat{y}_t)^2$	$\sqrt{(N_t - \hat{N}_t)^2}$
26	.0044	42
27	.0045	42
28	.0045	41

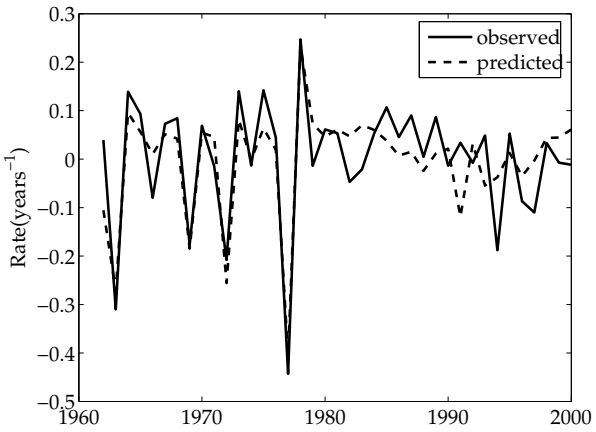
- LOO confirms:
  - the improvement over the previously considered models
  - that no great differences arise from models 26, 27, 28

- Hence we finally choose model 26, best according to SRM

Simulations with model 26



(a) Population abundances



(b) Rates of increase

Figure 2: Leave-one-out cross validation of model 26, best according to SRM.

Conclusions

- Our experiments show that SRM recognizes all the considered models with higher frequency than both FPE and SIC under practically all the simulation settings.

- A strength of SRM is its performance on small datasets and...

- the achievement of lower prediction errors in out-of-sample validation

- In the re-analyis of Alpine ibex case study, we add some further candidate models, achieving a general performances improvement of the SRM scores.

- One of these new models turns out to be the best selection, as confirmed also by leave-one-out cross validation.

Possibilities for future works

- Estimation of VC-dimensions of the commonest nonlinear models.

- Application of SRM and ICs in local modelling (lazy learning!)