

Erratum to “Model selection in demographic time series using VC-bounds” [Ecol. Modell. 191(1) (2006) 186-195]*

Giorgio Corani and Marino Gatto

Dipartimento Elettronica ed Informazione, Politecnico di Milano

e-mail: corani@elet.polimi.it

In the above publication we experimentally inter-compared various model selection criteria; we eventually recommended the use of the Structural Risk Minimization approach (SRM), as it showed consistently better performances than other approaches.

One of the considered criteria was the Schwarz Information Criterion (SIC). We adopted the formula of SIC quoted in several credited publications (Cherkassky and Mulier, 1998; Cherkassky et al., 1999), which is:

$$SIC = R_{emp}[1 + \frac{\ln(n)}{2}p(1-p)^{-1}] \quad (1)$$

where R_{emp} is the *empirical risk*, i.e., the mean square error measured on the model calibration dataset, n the number of samples in the dataset, p the parameters to data ratio (d/n). This formula is, however, wrong: as we have shown in (Corani and Gatto, 2006), it was erroneously derived from the formula originally provided in Schwarz (1978). Therefore, we have to warn against the use of formula (1) in future works dealing with SIC.

To correctly use SIC, one can start from the original formula proposed in Schwarz (1978):

$$SIC = -2 \ln(L) + d \ln(n) \quad (2)$$

where L is the likelihood function. With a few assumptions (see for instance Hastie et al. (2001)), an equivalent formulation in terms of empirical risk can be derived:

$$SIC = n \ln(R_{emp}) + d \ln(n) \quad (3)$$

In (Corani and Gatto, 2006) we experimentally inter-compare the performance of the wrong SIC (formula 1), the correct SIC (formula 3) and SRM over a variety of regression tasks. SIC performance largely improves by using the correct formula, as it selects more parsimonious models, is less affected by sample variability and leads to lower errors in out-of-sample predictions. However, our results show that SRM is anyway a more powerful model selection approach than SIC, in particular for small samples, which is often the case in ecology. So, we can confirm the original message of the paper, recommending SRM as the model selection approach to be preferred.

* Article history: Published online 22 September 2005
DOI of original article:10.1016/j.ecolmodel.2005.08.019

References

- Cherkassky, V., Mulier, F., 1998. Learning from data. Wiley-Interscience.
- Cherkassky, V., Shao, X., Mulier, F., Vapnik, V., 1999. Model complexity control for regression using VC generalization bounds. *IEEE Trans. on Neural Networks* 10 (5), 1075–1089.
- Corani, G., Gatto, M., 2006. Comments on: "Model complexity control for regression using VC generalization bounds: *IEEE TNN*, 10(5), 1999". Internal Report 2006.54, Dipartimento Elettronica ed Informazione, Politecnico di Milano.
URL www.elet.polimi.it/upload/corani/report2006-54.pdf
- Hastie, T., Tibshirani, R., Friedman, J., 2001. Elements of Statistical Learning. Springer Verlag.
- Schwarz, G., 1978. Estimating the dimension of a model. *Ann. Stat.* 6, 461–464.