

POLITECNICO DI MILANO

Dipartimento di Elettronica ed Informazione



Comments on: “Model Complexity
Control for Regression Using VC
Generalization Bounds”: IEEE TNN,
10(5), 1999

Giorgio Corani and Marino Gatto

Rapporto Interno n. 2006.54

Comments on: “Model Complexity Control for Regression Using VC Generalization Bounds”: IEEE TNN, 10(5), 1999

Giorgio Corani, Marino Gatto

Dipartimento di Elettronica e Informazione

Politecnico di Milano

e-mail: corani@elet.polimi.it

Abstract

In [1], various model selection approaches were experimentally inter-compared; one of the considered model selection criteria was the Schwarz Information Criterion (SIC); however, SIC was incorrectly implemented. The same mistake has been repeated in other more recent papers.

Here, we show why the SIC formula originally employed was wrong. We report instead the correct approach, which is well-known in statistics literature. We then show that the SIC performance is far better than the one described in [1], by repeating several experiments of the original paper. Nevertheless, we confirm that VC-based model selection is more powerful than SIC, especially for small samples.

1 Introduction

In [1] an empirical inter-comparison between different model selection criteria was carried out; particular emphasis was given to the model selection approach developed within the Statistical Learning Theory, and based on VC-dimension. In the following, such an approach will be referred to as VM (VC-based Model selection).

The model selection criteria considered in [1] included the Schwarz Information Criterion (SIC); however, the formula used for SIC was wrong. The same formula has been unfortunately employed in more recent papers [2, 3]. The present comment points out the mistake and examines how results vary if the correct approach is adopted.

In Section 2, we briefly analyze the Schwarz Information Criterion, reporting the correct formula to be used and showing why the SIC formula employed in [1] is wrong; in Section 3, we present a few experimental results, showing that the SIC fares much better than previously stated in [1], although its overall performance is anyway inferior to that of VM; in Section 4, we present the conclusions.

2 The Schwarz Information Criterion

The Schwarz Information Criterion [4] is designed to find the most probable model based on the available data: given a dataset of size n , a fixed set of models M_1, M_2, \dots, M_r and their posterior probabilities p_1, p_2, \dots, p_r (based on prior

probabilities and the likelihood of data), SIC seeks the model M_j such that, for $n \rightarrow \infty$, $p_j \rightarrow 1$ and $p_{\neq j} \rightarrow 0$ for all the models others than M_j [5]. Being derived within a Bayesian framework, it is often referred to also as Bayesian Information Criterion (BIC); an extensive discussion about Bayesian model selection and BIC can be found in [6].

2.1 The correct SIC formula

The SIC formula quoted in the original paper by Schwarz [4] reads as:

$$SIC = -2 \ln(L) + d \ln(n) \quad (1)$$

where L is the likelihood function, d the number of free parameters of the model, and n the number of samples in the dataset. Models are ranked according to increasing SIC.

Now consider input-output models of the kind $y = f(x, \theta) + \epsilon$, where x is the input, y the output, θ the parameter vector and ϵ the noise. Consistently with [1], let us denote the *empirical risk*, i.e., the error measured on the training set, as R_{emp} :

$$R_{emp} = \frac{1}{n} \sum_{i=1}^n (y_i - f(\mathbf{x}_i, \theta))^2 \quad (2)$$

Assuming that the noise is homoskedastic and normal, and denoting the noise variance as σ^2 , we have:

$$L = (\sigma^2 2\pi)^{-n/2} \exp\left(-\frac{n}{2\sigma^2} R_{emp}\right) \quad (3)$$

The log-likelihood is hence:

$$\ln(L) = -\frac{n}{2} \ln(\sigma^2) - \frac{n}{2} \ln(2\pi) - \frac{n}{2\sigma^2} R_{emp} \quad (4)$$

Dropping the constant term $-\frac{n}{2} \ln(2\pi)$, we have:

$$SIC = n \ln(\sigma^2) + \frac{n R_{emp}}{\sigma^2} + d \ln(n) \quad (5)$$

In most cases the variance of noise is unknown. Replacing σ^2 in expression 5 with its maximum likelihood estimate, i.e., $\hat{\sigma}^2 = R_{emp}$, and then dropping the constant term n , we obtain:

$$SIC = n \ln(\hat{\sigma}^2) + d \ln(n) = n \ln(R_{emp}) + d \ln(n) \quad (6)$$

Formula 6 is well-known in statistics [7, 8].

The number of free parameters d includes the noise variance, in addition to the parameters of the approximating function [5].

The practical use of SIC shows that simply choosing the model with the lowest SIC is not the best strategy. Instead, the application of a “rule of thumb” is recommended, which prescribes the consideration of the subset of models having

$\Delta SIC < 2$ with respect to the model with the lowest SIC [9]. The parsimony principle then suggests that the model with lowest d in the subset should be chosen.

In the experiments shown in Section 3, we will denote as SIC the simple-minded use of formula (6) (i.e., select the model with the lowest SIC) and as SIC* the use of formula (6) jointly with the rule of thumb and the principle of parsimony.

2.2 Analysis of the formula used in [1]

The formula for SIC given in [1, 2, 3] is

$$SIC = R_{emp} \left[1 + \frac{\ln(n)}{2} p(1-p)^{-1} \right] \quad (7)$$

We will now show that it is wrong. Also, we can somehow reconstruct how it was erroneously derived. Assuming that the variance σ^2 is known and equal for all the models, one can drop the first term in the right-hand side of equation 5, to obtain the formula quoted in [8]:

$$SIC = \frac{n}{\sigma^2} \left(R_{emp} + d \ln(n) \frac{\sigma^2}{n} \right) \quad (8)$$

Then, dividing by $\frac{n}{\sigma^2}$, we have:

$$SIC = R_{emp} + d \ln(n) \frac{\sigma^2}{n} \quad (9)$$

However, one still needs σ^2 to compute the SIC for each model. A possibly unbiased estimate of σ^2 is:

$$\sigma^2 = \frac{1}{n-d} \sum_{i=1}^n (y_i - f(\mathbf{x}_i, \theta))^2 = \frac{n}{n-d} R_{emp} \quad (10)$$

Now, both R_{emp} and d do depend on the model complexity; hence, there are as many σ^2 estimates as there are candidate models, which, by the way, contradicts the assumptions on σ^2 .

Disregarding this contradiction and substituting the estimated variance of formula (10) into the SIC expression (9), one obtains after some computations:

$$SIC = R_{emp} [1 + \ln(n) p(1-p)^{-1}] \quad (11)$$

where $p = d/n$. It still differs by a factor 1/2 from the formula published in [1], which indeed contains also a computational mistake in the final derivation (see formula 7 vs 11).

We now want to understand how much the results of [1], in which SIC is compared to VM, are sensitive to the incorrect implementation of SIC. In the experiments shown in Section 3, we will denote as SC the model selection criterion based on formula 7, which was actually used in [1]. As done in [1], we will not apply any rule of thumb or parsimony principle to SC.

3 Experimental results

In this section we empirically inter-compare SIC, SIC*, SC and VM; to this purpose, we retain the same experimental methodology designed in [1]. In particular, we use the publicly available source code ¹ of the Authors, modifying it only to incorporate SIC and SIC*; then, we repeat some of the experiments of the original paper. The VM criterion is largely described in [1], so we do not provide any detail about it here.

A brief description of the experimental methodology is provided in the following. Let us define the signal-to-noise ratio (SNR) as the ratio of the standard deviation of the true y (computed without noise) to the standard deviation of the noise, and let us recall that n denotes the training set size. We consider two different settings: ($n = 30$, $SNR = 2.5$) and ($n = 100$, $SNR = 2.5$).

For each setting, 300 training sets are artificially generated as follows: random x -values are drawn from a uniform probability distribution in the $[0,1]$ interval; y values are computed as $y = \sin^2(2\pi x) + \epsilon$, where ϵ is an additive Gaussian noise.

Two sets of experiments are performed, considering two different kinds of approximating functions:

- algebraic polynomials with m parameters:

$$f(x, \mathbf{w}) = \sum_{i=0}^{m-1} w_0 + w_i x^i \quad (12)$$

- trigonometric polynomials with m parameters:

$$f(x, \mathbf{w}) = \sum_{i=0}^{m-1} w_0 + w_i \cos(ix) \quad (13)$$

For both algebraic and trigonometric polynomials, we consider a set of candidates with complexity comprised between $m = 1$ and $m = 25$. Therefore, either 25 algebraic or 25 trigonometric polynomials are fitted to each generated dataset, and then used as candidates from which the “best” model must be chosen.

For each dataset, best models are chosen according to SIC, SIC*, SC, VC. Then, the prediction risk of the chosen models is empirically estimated by computing the mean square error on a 1000-samples testing set, generated analogously to the training set. This risk evaluation technique and its implementation are actually identical to that employed in [1].

Results are presented in Figure 1; the boxplots represent the empirical distributions, obtained from 300 experiments, of the prediction risk and the complexity m of the models chosen via the different criteria.

The results for $n = 30$ can be compared to those of Fig. 3 in [1]. One can thus verify that the performances of both VM and SC as reported here do coincide with those of [1].

However, let us now compare SC with the correct SIC and SIC*. Under every setting, SC chooses models more parameterized than SIC or SIC*; moreover, the choice of model complexity is much more affected by sample variability; the boxplots of m are indeed much wider than those of SIC or SIC*. Remarkably, even for $n = 100$, when SIC and SIC* do stabilize their choices around a restricted range of model complexities, SC continues to choose models of largely

¹www.ece.umn.edu/groups/ece8591/software/penal.zip

different complexities within the same set of experiments. Also, it is noteworthy that for $n = 100$ SIC, SIC* and VM converge to similar distribution of both m and the prediction risk, while SC does not; it consistently chooses more complex models. Correspondingly, the risk levels resulting from the implementation of SC are higher than those resulting from SIC or SIC*.

The use of SIC* should be preferred over SIC, as it is less affected by the random sample variability and generally leads to lower prediction risk. As the sample size increases, the difference between SIC and SIC* becomes smaller, and for $n = 100$ there is no critical difference between the two methods.

Despite SIC was wrongly implemented in [1], our results confirm that VM is a more powerful model selection approach than SIC, as already stated in [1]; in particular, for small samples it is the criterion most robust to sample variability (it provides the narrowest boxplots of both model complexity and prediction risk) and leads to remarkably lower average prediction risk than the SIC criteria. The difference is indeed important for small datasets, but not as large as claimed in [1]. For very large datasets the difference between the performances of VM and SIC becomes negligible.

4 Conclusions

Future works involving the use of SIC as model selection criterion should avoid the use of the SC formula quoted in [1]; it is wrong and performs very poorly in model selection. Instead, the correct formula (eq. 6 of this paper) provides much better performances than previously claimed. Also, we recommend the use of SIC jointly with the rule of thumb and the parsimony principle: select the model with lowest complexity within the subset of models having $\Delta SIC < 2$ with respect to the model with the lowest SIC. It leads to better results than the “simple-minded” use of SIC, namely, select the model with the lowest SIC.

In any case, the approach based on Statistical Learning Theory (VM) is a more powerful model selection method than SIC and even SIC*. The difference is really critical for samples of few tens of data, less important for larger datasets.

Acknowledgments

The work was partially supported by Istituto di Ingegneria Biomedica, CNR, Italy.

References

- [1] 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.
- [2] V. Cherkassky, “Model complexity control and statistical learning theory,” *Natural Computing*, vol. 1, no. 1, pp. 109–133, 2002.
- [3] H. Wechsler, Z. Duric, L. Fayin, and V. Cherkassky, “Motion estimation using statistical learning theory,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 4, pp. 466 – 478, 2004.

- [4] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, pp. 461–464, 1978.
- [5] K. Burnham and D. Anderson, "Multimodel inference - understanding AIC and BIC in model selection," *Sociological Methods & Research*, vol. 33, no. 2, pp. 261–304, 2004.
- [6] L. Wasserman, "Bayesian model selection and model averaging," *J. Math. Psychology*, vol. 44, no. 1, pp. 92–107, 2000.
- [7] M. Taper, "Model identification from many candidates," in *The nature of scientific evidence* (M. L. Taper and S. R. Lele, eds.), University of Chicago Press, 2004.
- [8] T. Hastie, R. Tibshirani, and J. Friedman, *Elements of Statistical Learning*. Springer Verlag, 2001.
- [9] A. Raftery, "Bayesian model selection in social research (with discussion)," *Sociological Methodology*, vol. 25, pp. 111–196, 1995.

Figure Captions

Figure 1. Results of the application of SIC, SIC*, SC and VM in different experimental settings. The boxes display the lower quartile, the median and the upper quartile, while the whiskers show the 5th and 95th percentiles.

Figures

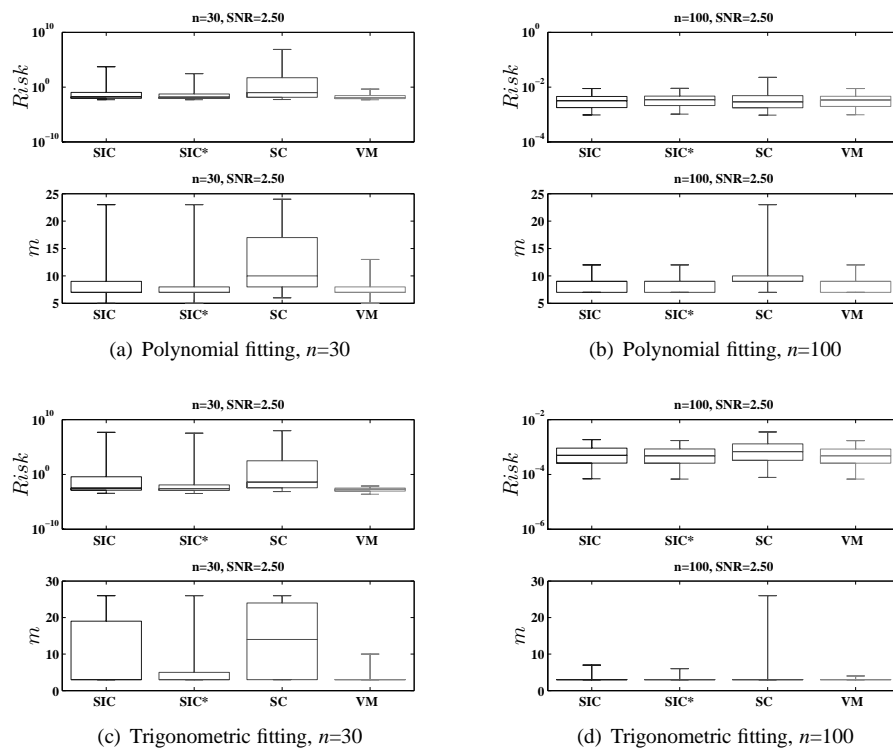


Figure 1: Results of the application of SIC, SIC*, SC and VM in different experimental settings. The boxes display the lower quartile, the median and the upper quartile, while the whiskers show the 5th and 95th percentiles.