MODEL SELECTION

Walter Zucchini [1]
Institute for Statistics and Econometrics, Georg-August-Universität
Platz der Göttinger Sieben 5, 37073 Göttingen, Germany

Gerda Claeskens [2]
ORSTAT and Leuven Statistics Research Center
K.U.Leuven, Naamsestraat 69, 3000 Leuven, Belgium

Georges Nguefack-Tsague [3]
Biostatistics Unit, Department of Public Heath
University of Yaoundé I, P. O. Box 1364 Yaoundé, Cameroon

1 Introduction

In applications there are usually several models for describing a population from a given sample of observations and one is thus confronted with the problem of model selection. For example, different distributions can be fitted to a given sample of univariate observations; in polynomial regression one has to decide which degree of the polynomial to use; in multivariate regression one has to select which covariates to include in the model; in fitting an autoregressive model to a stationary time series one must choose which order to use.

When the models under consideration is nested, as is the case in polynomial regression, the fit of the model to the sample improves as the complexity of the model (e.g. the number of parameters) increases but, at some stage, its fit to the population deteriorates. That is because the model increasingly moulds itself to the features of the sample rather than to the 'true model', namely the one that characterises the population. The same tendency occurs even if the models are not nested; increasing the complexity eventually leads to deterioration. Thus model selection needs to take both goodness of the fit and the complexity of the competing models into account.

Reference books on model selection include Linhart and Zucchini (1986), Burnham and Anderson (2002), Miller (2002), Claeskens and Hjort (2008). An introductory article is Zucchini (2000).

2 Information criteria – frequentist approach

The set of models considered for selection can be thought of as approximating models which, in general, will differ from the true model. The answer to the question 'Which approximation is best?' depends, of course, on how we decide to measure the quality of the fit. Using the Kullback-Leibler distance for this

[1]Email: wzucchi@uni-goettingen.de
[2]Email: Gerda.Claeskens@econ.kuleuven.be
[3]Email: gnguefack@yahoo.fr

leads to the popular Akaike Information Criterion (AIC, Akaike, 1973):

$$\mathrm{AIC}(M) = 2\log(L(\hat{\theta})) - 2p\,,$$

where M is the model, L the likelihood, and $\hat{\theta}$ the maximum likelihood estimator of the vector of the model's p parameters. The first term of the AIC measures the fit of the model to the *observed sample*; the fit improves as the number of parameters in the model is increased. But improving the fit of the model to the sample does not necessarily improve its fit to the population. The second term is a penalty term that compensates for the complexity of the model. One selects the model that maximizes the AIC. Note, however, that in much of the literature the AIC is defined as minus the above expression, in which case one selects the model that minimizes it.

A *model selection criterion* is a formula that allows one to compare models. As is the case with the AIC, such criteria generally comprise two components: one that quantifies the fit to the data, and one that penalizes complexity. Examples include Mallows' C_p criterion for use in linear regression models, Takeuchi's model-robust information criterion TIC, and refinements of the AIC such as the 'corrected AIC' for selection in linear regression and autoregressive time series models, the network information criterion NIC, which is a version of AIC that can be applied to model selection in neural networks, and the generalized information criterion GIC for use with influence functions. Several of these criteria have versions that are applicable in situations where there are outlying observations, leading to robust model selection criteria; other extensions can deal with missing observations.

Alternative related approaches to model selection that do not take the form of an information criterion are *bootstrap* (see, e.g., Zucchini, 2000) and *cross-validation*. For the latter the idea is to partition the sample in two parts: the calibration set, that is used to fit the model, and the validation sample, that is used to assess the fit of the model, or the accuracy of its predictions. The popular 'leave-one-out cross-validation' uses only one observation in the validation set, but each observation has a turn at comprising the validation set. In a model selection context, we select the model that gives the best results (smallest estimation or prediction error) averaged over the validation sets. As this approach can be computationally demanding, suggestions have been made to reduce the computational load. In 'five-fold cross-validation' the sample is randomly split in five parts of about equal size. One of the five parts is used as validation set and the other four parts as the calibration set. The process is repeated until each of the five sets is used as validation set.

3 Bayesian approach

The Bayesian regards the models available for selection as candidate models rather than approximating models; each of them has the potential of being the true model. One begins by assigning to each of them a prior probability, $P(M)$,

that it is the true model and then, using Bayes' theorem, computes the posterior probability of it being so:

$$P(M|\,\text{Data}) = \frac{P(\,\text{Data}\,|M)P(M)}{P(\,\text{Data}\,)}.$$

The model with the highest posterior probability is selected. The computation of $P(\,\text{Data}\,|M)$ and $P(\,\text{Data}\,)$ can be very demanding and usually involves the use of Markov chain Monte Carlo (MCMC) methods because, among other things, one needs to 'integrate out' the distribution of the parameters of M (see e.g. Wasserman, 2000).

Under certain assumptions and approximations (in particular the Laplace approximation), and taking all candidate models as *a priori* equally likely to be true, this leads to the Bayesian Information Criterion (BIC), also known as the Schwarz criterion (Schwarz, 1978):

$$\text{BIC}(M) = 2\log(L(\hat{\theta})) - p\log(n)\,,$$

where n is the sample size and p the number of unknown parameters in the model. Note that although the BIC is based on an entirely different approach it differs from the AIC only in the penalty term.

The difference between the frequentist and Bayesian approaches can be summarized as follows. The former addresses the question 'Which model is best, in the sense of least wrong?' and the latter the question 'Which model is most likely to be true?'.

The Deviance Information Criterion (Spiegelhalter et al., 2002) is an alternative Bayesian method for model selection. While explicit formulae are often difficult to obtain, its computation is simple for situations where MCMC simulations are used to generate samples from a posterior distribution.

The principle of minimum description length (MDL) is also related to the BIC. This method tries to measure the complexity of the models and selects the model that is the least complex. The MDL tries to minimize the sum of the description length of the model, plus the description length of the data when fitted to the model. Miminizing the description length of the data corresponds to maximizing the log likelihood of the model. The description length of the model is not uniquely defined but, under certain assumptions, MDL reduces to BIC, though this does not hold in general (Rissanen, 1996). Other versions of MDL come closer to approximating the full Bayesian posterior $P(M|\,\text{Data})$. See Grünwald (2007) for more details.

4 Selecting a selection criterion

Different selection criteria often lead to different selections. There is no clear-cut answer to the question of which criterion should be used. Some practitioners stick to a single criterion; others take account of the orderings indicated by two or three different criteria (e.g. AIC and BIC) and then select the one that leads

to the model which seems most plausible, interpretable or simply convenient in the context of the application.

An alternative approach is to tailor the criterion to the particular objectives of the study, i.e. to construct it in such a way that selection favours the model that best estimates the quantity of interest. The Focussed Information Criterion (FIC, Claeskens and Hjort, 2003) is designed to do this; it is based on the premise that a good estimator has a small mean squared error (MSE). The FIC is constructed as an estimator of the MSE of the estimator of the quantity of interest. The model with the smallest value of the FIC is the best.

Issues such as consistency and efficiency can also play a role in the decision regarding which criterion to use. An information criterion is called *consistent* if it is able to select the true model from the candidate models, as the sample size tends to infinity. In a weak version, this holds with probability tending to one; for strong consistency, the selection of the true model is almost surely. It is important to realize that the notion of consistency only makes sense in situations where one can assume that the true model belongs to the set of models available for selection. Thus will not be the case in situations in which researchers "believe that the system they study is infinitely complicated, or there is no way to measure all the important variables" (McQuarrie and Tsai, 1998). The BIC is a consistent criterion, as is the Hannan-Quinn criterion that uses $\log\log(n)$ instead of $\log(n)$ in the penalty term.

An information criterion is called *efficient* if the ratio of the expected mean squared error (or expected prediction error) under the selected model and the expected mean squared error (or expected prediction error) under its theoretical minimizer converges to one in probability. For a study of the efficiency of a model selection criterion, we do not need to make the assumption that the true model is one of the models in the search list. The AIC, corrected AIC, and Mallows's C_p are examples of efficient criteria. It can be shown that the BIC and the Hannan-Quinn criterion are not efficient. This is an observation that holds in general: consistency and efficiency cannot occur together.

5 Model selection in high dimensional models

In some applications, e.g. in radiology and biomedical imaging, the number of unknown parameters in the model is larger than the sample size, and so classical model selection procedures (e.g. AIC, BIC) fail because the parameters cannot be estimated using the method of maximum likelihood. For these so-called high-dimensional models regularized or penalized methods have been suggested in the literature. The popular Lasso estimator, introduced by Tibshirani (1996), adds an l_1 penalty for the coefficients in the estimation process. This has as a particular advantage that it not only can shrink the coefficients towards zero, but also sets some parameters equal to zero, which corresponds to variable selection. Several extensions to the basic Lasso exist, and theoretical properties include consistency under certain conditions. The Dantzig selector (Candes and Tao, 2008) is another type of method for use with high-dimensional models.

6 Post-model selection inference

Estimators that are obtained in a model that has been selected by means of a model selection procedure, are referred to as *estimators-post-selection* or *post-model-selection estimators*. Since the data are used to select the model, the selected model that one works with, is random. This is the main cause of inferences to be wrong when ignoring model selection and pretending that the selected model had been given beforehand. For example, by ignoring the fact that model selection has taken place, the estimated variance of an estimator is likely to be too small, and confidence and prediction intervals are likely to be too narrow. Literature on this topic includes Pötscher (1991), Hjort and Claeskens (2003), Shen et al. (2004), Leeb and Pötscher (2005).

Model selection can be regarded as the special case of model averaging in which the selected model takes on the weight one and all other models have weight zero. However, regarding it as such does not solve the problem because selection depends on the data, and so the weights in the estimator-post-selection are random. This results in non-normal limiting distributions of estimators-post-selection, and requires adjusted inference techniques to take the randomness of the model selection process into account. The problem of correct post-model selection inference has yet to be solved.

REFERENCES

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory.* eds. Petrov B. and Csáki F., Akadémiai Kiadó, Budapest, 267–281.

[2] Burnham, P. K. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach* (2nd edition). Springer-Verlag, New York.

[3] Candes, E. and Tao, T. (2008). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, **35**, 2313–2351.

[4] Claeskens, G. and Hjort, N. L. (2003). The focussed information criterion (with discussion). *Journal of the American Statistical Association*, **98**, 900–916.

[5] Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging.* Cambridge University Press, Cambridge.

[6] Grünwald, P. (2007). *The Minimum Description Length Principle*, MIT Press, Boston.

[7] Hjort, N. L. and Claeskens, G. (2003). Frequentist Model Average Estimators (with discussion). *Journal of the American Statistical Association*, **98**, 879–899.

[8] Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Fact and fiction. *Econometric Theory*, **21**, 21–59.

[9] Linhart, H. and Zucchini, W. (1986). *Model Selection.* Wiley, New York.

[10] McQuarrie, A. D. R and Tsai, C. L. (1998). *Regression and time series model selection.* World Scientific Publishing, River Edge.

[11] Miller, A. J. (2002). *Subset selection in regression* (2nd edition). Chapman and Hall/CRC, Boca Raton, Florida.

[12] Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, **7**, 163–185.

[13] Rissanen, J.J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, **42**, 40–47.

[14] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

[15] Shen, X., Huang, H. C. and Ye, J. (2004). Inference after model selection. *Journal of the American Statistical Association*, **99**, 751–762.

[16] Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society.* Series B, **64**, 583–639.

[17] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.* Series B, **58**(1), 267–288.

[18] Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, **44**, 92–107.

[19] Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, **44**, 41–61.

Reprinted with permission from Lovric, Miodrag (2011), International Encyclopedia of Statistical Science. Heidelberg: Springer Science+Business Media, LLC