On model selection from a finite family of possibly misspecified models

A Misspecification-Resistant Information Criterion

Hsiang-Ling Hsu, Ching-Kang Ing, and Howell Tong

Abstract Model selection problems are usually classified into two categories according to whether the data generating process (DGP) is included among the family of candidate models. The first category assumes that the DGP belongs to the candidate family, and the objective of model selection is simply to choose this DGP.The second category assumes that the DGP is not one of the candidate models. In this case, one of the top concerns is to choose the model having the best prediction capability. However, most existing model selection criteria can only perform well in at most one category, and hence when the underlying category is unknown, the choice of selection criteria becomes a key point of contention. In this article, we propose a misspecification-resistant information criterion (MRIC) to rectify this difficulty under the fixed-dimensional framework, which requires that the set of candidate models is fixed with the sample size. We prove the asymptotic efficiency of MRIC regardless of whether the true model belongs to the candidate family or not. We also illustrate MRIC's finite-sample performance using Monte Carlo simulation.

Key words: model selection, prediction error, misspecification-resistant information criterion

Consider finite parametric models. In many practical situations, we are often faced with the fundamental problem of selecting a model from a finite family of candidate models, none of which is necessarily the true data generating process, (DGP). Although existing literature on model selection is quite vast, the above problem does

Hsiang-Ling Hsu
Institute of Statistics, National University of Kaohsiung, Kaohsiung 811, Taiwan, e-mail: hsuhl@nuk.edu.tw;

Ching-Kang Ing
Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan, e-mail: cking@stat.sinica.edu.tw;

Howell Tong
University of Electronic Science and Technology, Chengdu, China; London School of Economics, UK, e-mail: howell.tong@gmail.com;

not seem to have received much attention. In fact, model selection problems are usually classified into two categories according to whether the true DGP is included among the family of candidate models. The first category assumes that the true DGP belongs to the candidate family, and the objective of model selection is simply to choose this DGP. The second category assumes that the true DGP is not one of the candidate models. In this case, one primary objective is to choose the model that has the best prediction capability. However, most existing model selection criteria can only perform well in at most one category, and hence when the underlying category is unknown, the choice of selection criteria becomes a serious point of contention. More seriously, none of them has addressed the fundamental problem mentioned at the opening sentence. In this article, we propose a misspecification-resistant information criterion (MRIC) to overcome this difficulty under the fixed-dimensional framework, which requires that the family of candidate finite-parametric models is fixed, independent of the sample size.

We prove the asymptotic efficiency of MRIC regardless of whether the true DGP belongs to the candidate family or not. We also illustrate MRIC's finite-sample performance using Monte Carlo simulation. Let us consider finite parametric models. Let us label the category as category I when the true DGP belongs to the candidate family. A model selection criterion is said to be consistent if it can choose the (most parsimonious) true DGP with probability tending to 1. In linear regression or time series models, Bayesian information criterion (BIC) (Schwarz 1978) has been shown to have this property; see, e.g., Nishii (1984), Rao and Wu (1989) and Wei (1992). On the other hand, Akaike's information criterion (AIC) (Akaike 1974) and Mallows' C_p (Mallows 1973), which tend to choose overfitting models, are not consistent in categorical I (e.g., Shibata 1976 and Shao 1997). The second category (category II) assumes that the true DGP is not one of the candidate models. In this category, choosing models having accurate prediction capabilities becomes the objective. When the true DGP is a linear regression model with infinitely many parameters and the number of predictor (explanatory) variables in the candidate models increases to infinity with the sample size, such that the corresponding approximation errors vanishes ultimately, Shibata (1981) and Li (1987) showed that AIC and Mallows' C_p possess asymptotic efficiency (AE), in the sense that these criteria can choose the model whose (finite-sample) mean squared prediction error (MSPE) is asymptotically equivalent to the smallest one among those of the candidate models. In contrast, BIC fails to achieve AE under category II; see Shibata (1980), Shao (1997) and Ing and Wei (2005). For a survey of the performance of various model selection criteria in both categories, see Shao (1997).

It is usually difficult for practitioners to perceive which category applies. As mentioned in the previous paragraph, most existing criteria cannot *simultaneously* enjoy consistency in category I and AE in category II. Consequently the choice of selection criteria has become a key point of contention over the past decade. Attempts have been made to address the contention. Ing (2007) and Yang (2007) have recently proposed similar adaptive procedures. They first compare two models selected by BIC, one for *partial* data points and other for *full* data points. They adopt AIC if the two selected models are different suggesting the plausibility of category II, and BIC otherwise. By suitably deciding the number of partial data points in the first step, they have shown that the proposed two-step procedure possesses consistency and AE in categories I and II, respectively. More recently, Liu and Yang (2011) devised the so called “parametricness index" to determine between categories I and II, and Zhang and Yang (2015) proposed using cross-validation to select between AIC and BIC in the absence of prior information on the underlying category. For a related result on solving the AIC-BIC dilemma from the point of view of cumulative risk, see van Erven et al. (2012). Although these recent efforts to resolve the controversy between AIC and BIC are novel, they mainly contribute to the increasing-dimensional (ID) framework, which allows the number of candidate predictor variables to grow to infinity with the sample size. The ID framework, however, may not be applicable to situations where collecting an increasing number of predictor variables is expensive, technically infeasible, or unnecessary according to domain knowledge (constraints). For instance, Kepler’s third law asserts that the ratio of the square of the revolutionary period to the cube of the orbital axis is the same for all the planets of the solar system. Therefore, if we wish to establish a statistical model for a planet’s period of revolution around the sun, predictor variables other than its orbital axis appear to be unessential, even when more data become available.

It can be argued that the really fundamental question is this: In many realistic situations, we are often faced with the problem of selecting a model from a *finite* family of candidate models, none of which is necessarily the true DGP. Although existing literature on model selection is quite vast, the above problem does not seem to have received much attention. This motivates us to ask whether there exists a model selection procedure that can perform well in both categories when the family of candidate models does not change with the sample size. We refer to this situation as the fixed-dimensional (FD) framework. It is already well known that when category I holds, BIC is consistent under both ID and FD frameworks; see Shao (1997) and Ing (2007). On the other hand, when category II holds instead of category I, AIC is AE under the ID framework but fails to carry over to the FD one. (Note that the definitions of AE in the FD and the ID frameworks are slightly different but similar in spirit.) Sin and White (1996) and Inoue and Kilian (2006) have shown that a BIC-type criteron has the so-call ‘strong parsimony property’ under the FD framework in the sense that it will asymptotically choose the most parsimonious model among those candidates having the smallest ‘population’ MSPE. However, as argued by Findley (1991), when two ‘misspecified’ models have the same population MSPE, the one with fewer parameters does not necessarily lead to the smaller (finite-sample) MSPE, which is the sum of the population MSPE and a term accounting for estimation error. This is in sharp contrast to the situation with two correctly specified models in which the smaller (finite-sample) MSPE is always given by the simpler model. As a result, AE is also not achievable by the BIC-type criteria under Category II and within the FD framework.

Indeed, there are already several criteria proposed to combat model misspecification, e.g., TIC (Takeuchi, 1976), GIC (Konishi and Kitagawa, 1996) and GBIC and GBIC$_p$ (Lv and Liu, 2014). However, it seems decidedly difficult to justify their AE under the FD framework. In this paper, we propose a misspecification-resistant information criterion (MRIC) to address arguably the most realistic situation in practice. Specifically, we prove that MRIC, within the FD framework, possesses AE regardless of whether the true DGP belongs to the candidate family or not. (Note that AE implies consistency under category I.) The MRIC has additional advantages. First, it is applicable to h-step prediction of time series data with $h \geq 1$. In particular, by changing the prediction lead times in the MRIC formula, the AE of MRIC is guaranteed for each $h \geq 1$. Second, unlike the ID conterparts of MRIC given in Ing (2007), Yang (2007) and Zhang and Yang (2015), MRIC can achieve AE on its own without the help of additional/auxiliary criteria. Third, by incorporating some screening methods, MRIC also performs satisfactorily in high-dimensional models. We summarize the performance of major model selection procedures discussed above in the form of the two tables; Table 1 is for the ID framework and Table 2 for the FD framework.

Table 1 Increasing-dimensional case (# of candidates increases with n)

Criteria	Case I: The model is included as a candidate Goal: Consistency	Case II: The model is NOT included as a candidate Goal: Asymp. efficiency for prediction (AE).	Case III: No info. on whether the true model is included Goal: Consistency when the true model is included + AE when the true model is not included.
AIC	No	Yes	No
BIC	Yes	No	No
GAIC	No	Yes	No
GBIC	Yes	No	No
Two-stage IC	Yes	Yes	Yes

Table 2 Fixed-dimensional case (# of candidates is fixed independent of n)

Criteria	Case I: consistency	Case II: AE	Case III: Consistency + AE
AIC	No	No	No
BIC	Yes	No	No
GAIC	No	No	No
GBIC	Yes	No	No
GBIC$_p$	Yes	No	No
MRIC	Yes	Yes	Yes

References

1. BOZDOGAN, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology* **44** 62–91.
2. BROCKWELL, P. J. and DAVIS, R. A. (1987) *Time series: theory and methods*. (1st ed.), Springer.

3. Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach* (2nd ed.), New York: Springer-Verlag.
4. Chan, N. H., and Ing, C.-K. (2011). Uniform moment bounds of fisher's information with applications to time series. *The Annals of Statistics* **39** 1526–1550.
5. Cleaskens, G., Croux, C., and Kerckhoven, J. V. (2007). Perdiction-focused model selection for autoregressive models. *Australian & New Zealand Journal of Statistics* **49** 359–379.
6. Davies, P.L. (2008). Approximating data (with discussion). *J. of Korean Stat. Soc* **37** 191-240.
7. Findley, D. F. (1991). Counterexamples to parsimony and BIC. *Annals of the Institute of Statistical Mathematics* **43** 505–514.

[Findley and Wei(1993)] Findley, D. F., and Wei, C. Z. (1993). Moment bounds for deriving time series CLT's and model selection procedures. *Statistica Sinica* **3** 453–480.

8. Findley, D. F., and Wei, C. Z. (2002). AIC, overfitting principles, and the boundedness of moments of inverse matrices for vector autoregressions and related models. *Journal of Multivariate Analysis* **83** 415–450.
9. Hansen, B. (2010). Multi-Step Forecast Model Selection. Manuscript.
10. Ing, C.-K. (2003). Multistep prediction in autoregressive processes. *Econometric Theory* **19** 254–279.
11. Ing, C.-K. (2004). Selecting optimal multistep predictors for autoregressive processes of unknown order. *The Annals of Statistics* **32** 693–722.
12. Ing, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Annals of Statistics* **35** 1238–1277.
13. Ing, C.-K., and Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica* **21** 1473–1513.
14. Ing, C.-K., and Wei, C. Z. (2003). On same-realization prediction in an infinite-order autoregressive process. *Journal of Multivariate Analysis* **85** 130–155.
15. Ing, C.-K., and Wei, C. Z. (2005). Order selection for same-realization predictions in autoregressive processes. *Annals of Statistics* **33** 2423–2474.
16. Inoue, A. and Kilian, L. (2006). On the selection of forecasting models. *Journal of Econometrics* **130** 273–306.
17. Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika* **83** 875–890.
18. Li, K. C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics* **15** 958–975.
19. Liu, W. and Yang, Y. (2011). Parametric or nonparametric? a parametricness index for model selection. *The Annals of Statistics* **39** 2074–2102.
20. Lv, J. and Liu, J. S. (2014). Model selection principles in misspecified models. *Journal of the Royal Statistical Society, Ser. B* to appear.
21. Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics* **12** 758–765.
22. Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76** 369–374.
23. Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica* **7** 221–264.
24. Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63** 117–126.
25. Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics* **8** 147–164.
26. Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.
27. Shumway, R. H., Azari, A. S. and Pawitan, Y. (1988). Modeling mortality uctuations in Los Angeles as functions of pollution and weather effects. *Environmental Research* **45** 224–241.
28. Shumway, R. H. and Stoffer, D. S. (2011). *Time series analysis and its applications: with R examples* (3rd ed.), New York: Springer.
29. Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics* **5** 595–620.

30. TAKEUCHI, K. (1976). The distribution of information statistic and the criterion of the adequacy of a model. *Suri-Kagaku (Mathematical Sciences)* **3** 12–18, (in Japanese).

31. WEI, C. Z. (1992). On predictive least squares principles. *The Annals of Statistics* **20** 1–42.

32. XIA, Y. and TONG, H. (2011) Feature matching (with discussion). *Statistical Science* **26** 21-46.

33. YANG, Y. (2007). Prediction/estimation with simple linear model: Is it really that simple? *Econometric Theory* **23** 1–36.