

PROCEEDINGS OF THE UNIVERSITY OF VAASA
DISCUSSION PAPERS 59

Seppo Pynnönen

A CONSISTENT CRITERION FOR ESTIMATING
THE DIMENSION OF A MODEL

Vaasa 1984

Contents

	Page
Abstract	2
1. Introduction	2
2. Akaike's information criterion	4
3. A consistent criterion for the dimension of a model	6
4. Asymptotic properties	8
5. Summary and conclusions	11
References	11

Abstract

Pynnönen, Seppo (1984). A consistent criterion for estimating the dimension of a model. Proceedings of the University of Vaasa. Discussion Papers 59, 12 p.

In this paper we consider the problem of selecting a suitable dimension of a model. We present the main theory of an information criterion introduced by Akaike (1973) for the dimensionality problem. We propose an asymptotically consistent criterion for dimensionality, based on Akaike's principles. Some asymptotic properties of these two criteria are also investigated.

Seppo Pynnönen, School of Business Studies, University of Vaasa, Raastuvankatu 31, SF-65100 Vaasa 10, Finland.

1. Introduction

In applied statistics we are often faced with the problem of selecting an appropriate dimension of a model. Examples of this are variable selection in regression analysis, number of factors in classic factor analysis, and choosing the length of lag in an autoregressive and related problems.

Selection of the dimension should at first hand be based on the theory of the subject. Thus, for example in econometrics theoretical considerations may give a clue for the dimension and also perhaps the sign of a parameter and its range of possible values. However, other things being equal, the decision of the dimension must sometimes be based on data at hand. So it is reasonable to try to develop test procedures and single statistics, on which we can base our decision about the appropriate dimension. A classic approach to this problem has been iterative estimation-test procedure, in which the parameters are first estimated for each model and then tested. This procedure is repeated until all the remaining estimates appear to be statistically significant. In certain cases this approach leads to logical contradictions. For example in regression analysis the F-statistic of the forward selection procedure is based on the idea that the model which has one parameter more than the model under null hypothesis is the right one if the null hypothesis is rejected. However, in the next step this alternative will be the new null hypothesis, so if we reject it at this step we have a logical contradiction with the previous step. It is also a methodological contradiction except the special case, where regressors are orthogonal. Another point for criticism that concerns hypothesis testing generally is the degree of subjectivity in choosing the significance level of a test.

As alternatives to this iterative estimating-testing procedure different criterion functions has been developed, on which one can base his decision about the order of a model. A good presentation of the most important of these methods can be found in Amemiya (1980) (see also Hocking (1976), Thompson (1978), and Judge et al. (1980) ch. 11).

In this paper we shall consider the problem of selecting the dimension as an estimation theoretic problem, to be understood as an extension of the method of maximum likelihood (ML) to multimodel situation. It is well known that the method of maximum likelihood as such is not applicable to this situation, because it invariably leads to the largest dimension, and so it does not respond to our intuitive concept of the right dimension of a model. Akaike (1973) has expressed such an extension to ML-method by suggesting the selection of that alternative which minimizes the criterion

$$(1) \quad \text{AIC}(k) := -2 \log L_k(\hat{\theta}) + 2k,$$

where

- log: natural logarithm
- L: the likelihood function
- $\hat{\theta}_k$: ML estimator of ${}_k\theta = (\theta_1, \dots, \theta_k, 0, \dots, 0)$
- k: the number of parameters to be estimated
- $k = 0, 1, \dots, K < \infty$.

We shall present the theoretic backgrounds of Akaike's information criterion in chapter 2. In chapter 3 we shall suggest a Schwartz (1978) type information criterion based on Akaike's point of view. As a result we get a Consistent Information Criterion

$$(2) \quad \text{CIC}(k) := -\log L({}_k\hat{\theta}) + (1/2)k \log(n/2\pi).$$

The decision strategy is similar to AIC, that is, select that alternative which minimises CIC. In chapter 4 we investigate some asymptotic properties of AIC and CIC.

2. Akaike's information criterion

In this chapter we follow - as far as it concerns AIC - Amemiya's (1980) article. Let's denote by $L(\theta, X)$ or as above briefly by $L(\theta)$ the likelihood function, where θ denotes K - dimensional parameter vector and X is a n dimensional random vector where the components are not necessarily independent. Let θ_0 denote the true parameter value and ${}_k\theta$ the restricted alternative, $k = 0, \dots, K < \infty$. Akaike (1973) assumes that ${}_k\theta_0$ and θ_0 are situated very near each other, where ${}_k\theta_0$ is defined by (4) such that $W(\theta_0, {}_k\theta_0) = \min_{{}_k\theta} W(\theta_0, {}_k\theta)$. In other words he assumes that θ_0 "almost" satisfies the restrictions introduced to ${}_k\theta$. We put this hypothesis explicitly, i.e. we assume that θ_0 satisfies the restrictions set by the hypothesis

$$(3) \quad H: \theta = {}_k\theta = (\theta_1, \dots, \theta_k, 0, \dots, 0).$$

Let $\hat{\theta}$ and ${}_k\hat{\theta}$ denote the maximum likelihood estimates of θ and ${}_k\theta$ respectively. For estimation Akaike suggest the loss function

$$(4) \quad W(\theta_o, \hat{\theta}) := -(2/n) \int \log \left(\frac{L(\hat{\theta}, x)}{L(\theta_o, x)} \right) L(\theta_o, x) dx,$$

where $\hat{\theta}$ is understood as constant under integration, and x is a n -vector. Thus W defines Kullback-Leibler's mean discrimination information multiplied by two. Hence $W(\theta_o, \hat{\theta}) > 0$ whenever $\theta_o \neq \hat{\theta}$, and equals to zero if and only if $\theta_o = \hat{\theta}$. This property is repeatedly exploited in the following proofs of the main results of this paper. Since θ_o is unknown Akaike suggest to estimate W by

$$(5) \quad K^{\Lambda}_k := -(2/n) \log \left(\frac{L(\hat{\theta})}{L(k^{\hat{\theta}})} \right).$$

Although (5) is a consistent estimator to W , it has the drawback that it attains its minimum trivially only, if we select k equal to K . Akaike proceeds as follows to find a more useful and yet accurate approximation to W .

By assuming usual regularity conditions (see e.g. Cox and Hinkley (1974), ch. 9), we get by using the Taylor series expansion with respect to $\hat{\theta}$ in the neighbourhood of k^{θ}

$$(6) \quad W(k^{\theta}, \hat{\theta}) = ||k^{\hat{\theta}} - k^{\theta}||^2 + O(n^{-3/2}),$$

where $||k^{\hat{\theta}} - k^{\theta}||^2 := -(1/n)(k^{\hat{\theta}} - k^{\theta})' E[D_{\theta}^2 \log L(k^{\theta}, X)](k^{\hat{\theta}} - k^{\theta})$, D_{θ} denotes differentiation with respect to θ , $'$ denotes transpose, and $O(n^{-3/2})$ is the remainder.

In (6) we have also exploited the fact that $E D_{\theta} \log L(k^{\theta}, X) = 0$ under the hypothesis (3). By expanding next $\log L(k^{\theta}, X)$ into the Taylor series with respect to k^{θ} around $k^{\hat{\theta}}$, and taking into account that $D_{\theta} \log L(k^{\hat{\theta}}, X) = 0$, we have

$$(7) \quad -(2/n) \log L(k^{\theta}, X) = -(2/n) \log L(k^{\hat{\theta}}, X) + ||k^{\hat{\theta}} - k^{\theta}||^2 + O(n^{-3/2})$$

Similarly we have

$$(8) \quad -(2/n) \log L(\theta_o, X) = -(2/n) \log L(k^{\hat{\theta}}, X) + ||k^{\hat{\theta}} - \theta_o||^2 + O(n^{-3/2}).$$

Under the hypothesis (3) the left hand sides of (7) and (8) are equal. Thus we get an approximation to K^{Λ}_k into the form

$$(9) \quad K^{\Lambda}_k = \|\|_K \hat{\theta} - \theta_0 \|\|^2 - \|\|_k \hat{\theta} - k\theta \|\|^2 + O(n^{-3/2}).$$

Hypothesis (3) being valid we get by (6) and by a chi-square approximation of $n \|\|_k \hat{\theta} - k\theta \|\|^2$ that $E W(\theta_0, k\hat{\theta}) \approx k/n$. Similarly $E \|\|_K \hat{\theta} - \theta_0 \|\|^2 \approx K/n$ and $E \|\|_k \hat{\theta} - k\theta \|\|^2 \approx k/n$. Thus

$$(11) \quad E(K^{\Lambda}_k - K/n + 2k/n) \approx E W.$$

By ignoring $O(n^{-1/2})$ and the constant terms, we get Akaike's information criterion (1).

Through these considerations we see instantly that there exist no trivial transformation for AIC to be a "good" estimator to W , when the hypothesis (3) is not valid, in the general case (i.e. when the Fisher information matrix $-E D_{\theta}^2 \log L(\theta, X)$ is nondiagonal). However, this appears not to be a serious drawback of this method, since as we shall see in chapter 4, the probability of underfitting a model diminishes asymptotically to zero (see also Shibata 1976, Bhansali and Downham 1977 and Geweke and Meese 1981). The same is not true with the probability of overfitting a model (see ch. 4).

To eliminate this inconsistency, we propose in the following chapter a different approximation to the log-likelihood ratio to get an information criterion that has the property of consistency.

3. A consistent criterion for the dimension of a model

To make considerations simpler we assume in this and the subsequent chapter that the components of X are independent and identically distributed. By assuming that θ_0 satisfies the restrictions of ${}_k \theta$ (i.e. the hypothesis (3) is in effect), ${}_k \hat{\theta}$ is also the maximum likelihood estimator of θ_0 . It also is, at least asymptotically, a sufficient statistic for θ in this case. In fact, by assuming the usual regularity conditions and hypothesis (3), we have $\theta_0 = {}_k \hat{\theta} + O(n^{-1/2})$,

and by simple manipulation of (7) we get

$$(12) \quad L(\theta_0, X) = g(\hat{\theta}_k; \theta_0)h(X),$$

where $g(\hat{\theta}_k; \theta_0) := \exp\{-(n/2) \|\hat{\theta}_k - \theta_0\|^2 + O(n^{-1/2})\}$ and $h(X) := L(\hat{\theta}_k; X)$, which does not depend on θ .

Thus under the assumptions above, $\hat{\theta}_k$ is at least asymptotically sufficient for θ .

On the other hand if the assumptions made are valid, the asymptotic distribution of $\hat{\theta}_k$ is $N(\theta_0, (nI_{\theta_0})^{-1})$, where I_{θ_0} is a $k \times k$ Fisher information matrix at θ_0 with respect to one observation. So we can approximate the likelihood of θ at θ_0 by

$$(13) \quad L^*(\theta_0, \hat{\theta}_k) = \frac{(n^k |I_{\theta_0}|)^{1/2}}{(2\pi)^{k/2}} \exp\{-(n/2)(\hat{\theta}_k - \theta_0)' I_{\theta_0} (\hat{\theta}_k - \theta_0)\} * \\ (1 + O(n^{-1/2})).$$

Hence by sufficiency, we can write

$$(14) \quad L(\theta_0, X) = c(X)L^*(\theta_0, \hat{\theta}_k),$$

where $c(X)$ does not depend on θ , and thus includes no information about it.

We suggest to replace in the log-likelihood ratio (5) the estimator $L(\hat{\theta}_k)$ of $L(\theta_0)$ by $L(\theta_0) = c(X)L^*(\theta_0, \hat{\theta}_k)$, approximating θ_0 by $\hat{\theta}_k$. Thus we get

$$(15) \quad \eta_k := -(2/n) \left[\log L(\hat{\theta}_k) - \log L(\theta_0) \right] \\ = -(2/n) \log L(\hat{\theta}_k) + (k/n) \log(n/2\pi) \\ + (1/n) \log(|I_{\theta_0}|) + (2/n) \log(c(X)) + O(n^{-3/2})$$

By ignoring the constant terms and $O(n^{-1/2})$, we get the criterion (2).

As in the case of AIC, we must emphasize that CIC is not transformable by any simple transformation to a useful estimator of W in the general case, when the hypothesis (3) is not valid.

4. Asymptotic properties

As we already noted in the previous chapter, we assume in this chapter also that the components of X are independent and identically distributed. Let us denote AIC and CIC generally,

$$(16) \quad IC(k) := -\log L_k(\hat{\theta}) + g(n)k.$$

Thus in AIC $g(n) = 1$ and in CIC $g(n) = (1/2)\log(n/2\pi)$. Let us denote the dimension estimated by IC,

$$(17) \quad \hat{k} := \min \{ k \mid IC(k) \leq IC(m), m = 0, \dots, K \}.$$

We shall first show that the probability of underfitting a model vanishes asymptotically.

In fact, if $k < k_0$, where k_0 is the right dimension of the model:

Then by definition of \hat{k} we have

$$\begin{aligned} P(\hat{k} = k) &= P \{ IC(k) \leq IC(m); m = 0, \dots, K \} \leq P \{ IC(k) \leq IC(k_0) \} = \\ &= P \left\{ -(1/n) \log \left(\frac{L_k(\hat{\theta})}{L_{k_0}(\hat{\theta})} \right) \leq \frac{g(n)}{n} (k_0 - k) \right\}. \end{aligned}$$

Now $k_0 \hat{\theta} \xrightarrow{P} \theta_0$, and by defining $k_0 \theta_0$ such that

$$(18) \quad W(\theta_0, k_0 \theta_0) = \inf_{k \theta_0} W(\theta_0, k \theta_0),$$

we have $k \hat{\theta} \xrightarrow{P} k_0 \theta_0$ (cf. Akaike 1973: 273).

Since

$$(19) \quad -\frac{1}{n} \log \left(\frac{L_{k_0}(\hat{\theta})}{L_k(\hat{\theta})} \right) \xrightarrow{P} I(\theta_{0,k}, \theta_0) > 0,$$

where $I(\theta_{0,k}, \theta_0)$ denotes Kullback-Leibler's information, and since $(k_0 - k)g(n)/n \rightarrow 0$, we have

$$P(\hat{k} = k) \leq P\left\{-\frac{1}{n} \log \left(\frac{L_{k_0}(\hat{\theta})}{L_k(\hat{\theta})} \right) \leq g(n)(k_0 - k)/n\right\} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Thus we have shown that the probability of underfitting a model tends to zero as n tends to infinity.

Next we shall show that the probability of overfitting a model disappears asymptotically for CIC.

Let $k \geq k_0$:

In a situation where the values of the criterion function are calculated for every alternative and the minimum of the criterion function is selected, we can assume by the previous considerations that all the $k_0 - 1$ parameters already existing in the model are different from zero (i.e. they really belong into the model). Thus for CIC, we get

$$P(\hat{k} = k_0) = P\{CIC(k_0) \leq CIC(m); m = k_0 + 1, \dots, K\}$$

$$= P\left\{-\log \left(\frac{L_{k_0}(\hat{\theta})}{L_m(\hat{\theta})} \right) \leq (1/2)(m - k_0) \log(n/2\pi); m = k_0 + 1, \dots, K\right\}.$$

Now $-\log\{L_{k_0}(\hat{\theta})/L_m(\hat{\theta})\} \geq 0$, and multiplied by two is asymptotically chi-squared distributed with degrees of freedom $m - k_0$. Thus

$$\begin{aligned}
P(\hat{k} = k_0) &= P \left\{ -2 \log \left(\frac{L(k_0^{\hat{\theta}})}{L(m^{\hat{\theta}})} \right) \leq (m - k_0) \log(n/2\pi) \right\} \\
&\geq 1 - \frac{m - k_0}{(m - k_0) \log(n/2\pi)} \rightarrow 1 \text{ as } n \text{ tends to } \infty.
\end{aligned}$$

Thus we have shown that CIC gives a consistent estimate to the true order of a model. AIC, however, overestimates the true dimension with positive probability. To see this, it suffices to restrict into a nested case. Indeed, if $k > k_0$, we have

$$P \{ \text{AIC}(k) < \text{AIC}(k_0) \} = P \left\{ 2 \log \left(\frac{L(k^{\hat{\theta}})}{L(k_0^{\hat{\theta}})} \right) > 2(k - k_0) \right\}.$$

Using again the chi-squared approximation and since $2(k - k_0) < \infty$, we can conclude that the probability above remains positive as n tends to infinity. Thus AIC does not give a consistent estimate to the true dimensionality.

Using the same reasoning as above, and assuming the usual regularity conditions, we get, $\hat{k}^{\hat{\theta}} \xrightarrow{P} \theta_0$, where $\hat{k}^{\hat{\theta}}$ and \hat{k} are the estimates of θ_0 and k_0 respectively given by IC.

In fact, if θ_0 does not satisfy the constraints asserted to k^{θ} , we have similarly as in (19),

$$(20) \quad -(1/n) \log \left(\frac{L(k^{\hat{\theta}})}{L(\theta_0)} \right) \xrightarrow{P} I(\theta_0, k^{\theta_0}) > 0,$$

where k^{θ_0} is defined as in (18). Hence inserting for clarity k^{θ} in place of k in IC, we have $P\{IC(k^{\hat{\theta}}) < IC(\hat{\theta}_0)\} \rightarrow 0$ as $n \rightarrow \infty$ because of (20) and the fact that $g(n)(k - k_0)/n \rightarrow 0$ as $n \rightarrow \infty$. Thus we can conclude that $k^{\hat{\theta}} \xrightarrow{P} \theta_0$.

Hence, in conclusion we have that both AIC and CIC lead to a consistent estimate for θ_0 , and CIC in addition gives a consistent estimate to the right dimension of θ_0 .

5. Summary and conclusions

In this paper we have considered Akaike's (1973) information criterion (AIC), and proposed a Schwarz (1978) -type information criterion, starting from Akaike's principles. We also have investigated some asymptotic properties of these criteria. By these considerations we have concluded that both of these criteria yield asymptotically a consistent estimate for the true parameter value θ_0 . In addition to this the criterion proposed by us, yields a consistent estimate to the true order of θ_0 . Finally we can note that because of our asymptotic approach in the considerations above, and the maximal dimensionality being finite, we could just leave the divisor 2π off from $\log(n/2\pi)$ of CIC, and hence be led to Schwarz criterion $SBIC(k) = -\log L_k(\hat{\theta}) + (1/2)k \log(n)$.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Proceedings of the 2nd International Symposium on Information Theory, eds. B.N. Petrov and F. Czaki, 267-81. Budapest: Akademiai Kiado.
- Amemiya, T. (1980). Selection of regressors. *International Economic Review* 21:2, 331-54.
- Bhansali, R.J. and Downham, D.Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE-criterion. *Biometrika* 64, 547-51.
- Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.
- Geweke, J. and Meese, R. (1981). Estimating regression of finite but unknown order. *International Economic Review* 22:1, 55-70.
- Hocking, R.R. (1976). The analysis and selection of variables in linear regression. *Biometrics* 32, 1-49.

- Judge, C.G., Griffiths, W.E. Hill, R.C. and Lee, T-C. (1980). *The Theory and Practice of Econometrics*. New York: John Wiley.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461-4.
- Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* 63, 117-26.
- Thompson, M.L. (1978). Selection of variables in multiple regression: Part I. A review and evaluation. *International Statistical Review* 46, 1-19.
- Thompson, M.L. (1978). Selection of variables in multiple regression: Part II. Cosen procedures, computations and examples. *International Statistical Review* 46, 129-46.

PROCEEDINGS OF THE UNIVERSITY OF VAASA

Discussion Papers

37. VESA ROUTAMAA. Private and public sector management: An approach to structural differences. 1981.
38. KAUKO MIKKONEN. Teollisuusalueiden muodostamisesta. 1981.
39. VESA ROUTAMAA. Automation and administrative personnel: An analysis of the associations between automation and the ratios of the different administrative groups within organizations. 1981.
40. VESA SAARIO. Asymptotic properties of a sequential decision process under uncertainty. 1982.
41. REIJO RUUHELA & TIMO SALMI & MARTTI LUOMA & ARTO LAAKKONEN. Direct estimation of the internal rate of return from published financial statements. 1982.
42. MARTTI LUOMA & MAURI PALOMÄKI. A new theoretical gravitation model and its application to a case with drastically changing mass component. 1982.
43. MARTTI LUOMA. A new approach to the analysis of the financial time series of a corporation: Unobservable variables in accounting and finance. 1982.
44. JAAKKO ASTOLA & ILKKA VIRTANEN. Entropy correlation coefficient, a measure of statistical dependence for categorized data. 1982.
45. HENRIK NIKULA. Übersetzungstheorie und Pragmatik. 1982.
46. VESA ROUTAMAA. A comparison of organizational structures in different countries. 1982.
47. RISTO VÄNTSI. Kuluttajansuojan periaatteiden kokonaispuitemalli. 1982.
48. RISTO VÄNTSI. On the express warranty of quality and the marketing channel. 1983.
49. RISTO VÄNTSI. On the status of the service contract. 1983.
50. RISTO VÄNTSI. Kuluttajien informaationtarve ja kuluttajainformaation tarjonta. 1983.
51. MARKKU HAKULI & VESA ROUTAMAA. Liikkeenjohdon työn ja tehtävien tutkimus ja eräs mahdollisuus sen suuntaamiseen. 1983.
52. PAAVO YLI-OLLI & JUHA KULPAKKO & JYRY TOLVANEN. The impact of inflation on portfolio selection: Empirical evidence on Finnish stock markets. 1983.
53. RISTO VÄNTSI. Sekundaarisesta käyttäytymisnormista sopimustilanteessa. 1983.
54. MARTTI LUOMA. Eräitä näkökohtia yrityksen kasvun estimoinnista lähinnä tilastotieteen näkökulmasta. 1983.
55. RISTO VÄNTSI. Alistetun urakan tarkastaminen. 1983.
56. HARRI HIETIKKO. Hypothesis testing in simulation: Application to the time series forecasting. 1983.

PROCEEDINGS OF THE UNIVERSITY OF VAASA

Discussion Papers 59

57. ARI SALMINEN & ANITA NIEMI-IILAHTI. Itsehallinnosta. Johdatusta kunnallisen itsehallintokäsitteen muotoutumiseen. 1983.
58. ARI SALMINEN & KARI KUOPPALA. Moderni julkisen talouden organisaatio hallintotieteellisenä tutkimuskohteena: syntetisoiva tarkastelu. 1984.
59. SEPPO PYNNÖNEN. A consistent criterion for estimating the dimension of a model. 1984.