Estimating the Dimension of a Model

Gideon Schwarz

Annals of Statistics, Volume 6, Issue 2 (Mar., 1978), 461-464.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28197803%296%3A2%3C461%3AETDOAM%3E2.0.CO%3B2-5

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Annals of Statistics is published by Institute of Mathematical Statistics. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ims.html.

Annals of Statistics
©1978 Institute of Mathematical Statistics

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2002 JSTOR

http://www.jstor.org/
Mon Sep 23 05:22:59 2002

The Annals of Statistics
1978, Vol. 6, No. 2, 461-464

ESTIMATING THE DIMENSION OF A MODEL[1]

BY GIDEON SCHWARZ

Hebrew University

The problem of selecting one of a number of models of different dimensions is treated by finding its Bayes solution, and evaluating the leading terms of its asymptotic expansion. These terms are a valid large-sample criterion beyond the Bayesian context, since they do not depend on the a priori distribution.

1. Introduction. Statisticians are often faced with the problem of choosing the appropriate dimensionality of a model that will fit a given set of observations. Typical examples of this problem are the choice of degree for a polynomial regression and the choice of order for a multi-step Markov chain.

In such cases the maximum likelihood principle invariably leads to choosing the highest possible dimension. Therefore it cannot be the right formalization of the intuitive notion of choosing the "right" dimension. An extension of the maximum likelihood principle is suggested by Akaike [1] for the slightly more general problem of choosing among different models with different numbers of parameters. His suggestion amounts to maximizing the likelihood function separately for each model j, obtaining, say, $M_j(X_1, \cdots, X_n)$, and then choosing the model for which $\log M_j(X_1, \cdots, X_n) - k_j$ is largest, where k_j is the dimension of the model. We present an alternative approach to the problem.

In a model of given dimension maximum likelihood estimators can be obtained as large-sample limits of the Bayes estimators for arbitrary nowhere vanishing a priori distributions.

Therefore we look for the appropriate modification of maximum likelihood for our case, by studying the asymptotic beavior of Bayes estimators under a special class of priors. These priors are not absolutely continuous, since they put positive probability on some lower-dimentional subspaces of the parameter space, namely the subspaces that correspond to the competing models. In the large-sample limit, the leading term of the Bayes estimator turns out to be just the maximum likelihood estimator. Only in the next term something new is obtained. This was to be expected, since (as was shown in [2] and [3], albeit for sequential testing) the leading term depends on the prior only through its support, while the second order term does reflect singularities of the a priori distribution. We shall arrive at the following procedure:

Choose the model for which $\log M_j(X_1, \cdots, X_n) - \frac{1}{2}k_j \log n$ *is largest.*

The validity of this procedure as a large-sample version of Bayes procedures

Received August 1976; revised February 1977.

[1] Written while the author was a Fellow of the Institute for Advanced Studies on Mt. Scopus.

AMS 1970 *subject classifications*. Primary 62F99, 62J99.

Key words and phrases. Dimension, Akaike information criterion, asymptotics.

will be established here for the case of independent, identically distributed observations, and linear models.

2. The exact Bayes procedure. In a general parameter space, there is no intrinsic linear structure. We therefore assume that observations come from a Koopman–Darmois family, i.e., relative to some fixed measure on the sample space they possess a density of the form

$$f(x, \boldsymbol{\theta}) = \exp(\boldsymbol{\theta} \cdot \mathbf{y}(x) - b(\boldsymbol{\theta})),$$

where $\boldsymbol{\theta}$ ranges over the natural parameter space Θ, a convex subset of the K-dimensional Euclidean space, and $\mathbf{y}$ is the sufficient K-dimensional statistic. The competing models are given by sets of the form $m_j \cap \Theta$, where each m_j is a k_j-dimensional linear submanifold of K-dimensional space.

Fitting the asymptotic nature of the result, the a priori distribution need not be known exactly. It suffices to assume that it is of the form $\sum \alpha_j \mu_j$, where α_j is the a priori probability of the jth model being the true one, and μ_j, the conditional a priori distribution of θ given the jth model, has a k_j-dimensional density that is bounded and locally bounded away form zero throughout $m_j \cap \Theta$. This implies mutual orthogonality of the μ_j, since the intersection of two distinct linear manifolds either is one of them, or has lower dimensions than both.

Finally, we assume a fixed penalty for guessing the wrong model. (Actually, a loss that depends on $\boldsymbol{\theta}$ and on the guess would yield the same asymptotic results, provided the loss function stays between two fixed positive bounds for all wrong decisions.) Under this assumption, the Bayes solution consists of selecting the model that is a posteriori most probable. Via Bayes' formula that is equivalent to choosing the j that maximizes

$$S(\mathbf{Y}, n, j) = \log \int \alpha_j \exp((\mathbf{Y} \circ \boldsymbol{\theta} - b(\boldsymbol{\theta}))n)\, d\mu_j(\boldsymbol{\theta}),$$

where the integral extends over $m_j \cap \Theta$, and $\mathbf{Y}$ is the averaged $\mathbf{y}$-statistic $(1/n) \sum \mathbf{y}(X_i)$.

3. Asymptotics. The asymptotic expansion of $S(\mathbf{y}, n, j)$ could be obtained from results in an earlier paper [3] as a special case. We shall, however, keep this paper self-contained by outlining a proof of the necessary result directly.

PROPOSITION. *For fixed* $\mathbf{Y}$ *and* j, *as* n *tends to* ∞,

$$S(\mathbf{Y}, n, j) = n \sup (\mathbf{Y} \circ \boldsymbol{\theta} - b(\boldsymbol{\theta})) - \tfrac{1}{2} k_j \log n + R$$

where the remainder $R = R(\mathbf{Y}, n, j)$ *is bounded in* n *for fixed* $\mathbf{Y}$ *and* j.

PROOF. We shall proceed in steps.

LEMMA 1. *The proposition holds when* $\mathbf{Y} \circ \boldsymbol{\theta} - b(\boldsymbol{\theta}) = A - \lambda \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$ *where* $\lambda > 0$, $\boldsymbol{\theta}_0$ *is a fixed vector in* m_j, *and* μ_j *is Lebesgue measure on* m_j.

Explicit evaluation of the integral yields $\alpha_j(\pi/n\lambda)^{k_j/2}e^{nA}$, and

$$\sup A - \lambda||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 = A\,.$$

Therefore

$$S(\mathbf{Y}, n, j) = nA - \tfrac{1}{2}k_j \log(n\lambda/\pi) + \log \alpha_j$$

establishes the proposition for this case, with $R = \frac{1}{2}k_j \log(\pi/\lambda) + \log \alpha_j$.

LEMMA 2. *If two bounded positive random variables U and V agree on the set where either exceeds ρ for some $0 < \rho < \sup U$, then*

$$\log E(U^n) - \log E(V^n) \to 0$$

as $n \to \infty$.

Clearly it suffices to show that this holds for V that vanishes where $U \leqq \rho$. In this case $0 \leqq U^n - V^n \leqq \rho^n$, and therefore

$$E(V^n) \leqq E(U^n) \leqq E(V^n) + \rho^n = E(V^n)\left(1 + \frac{\rho^n}{E(V^n)}\right)$$

and we only have to show $\log(1 + (\rho^n/E(V^n))) \to 0$. Now $(E(V^n))^{1/n} \to \sup V$ (a well-known fact on L_n norms) and $\sup V = \sup U > \rho$ yield for $\rho/(E(V^n))^{1/n}$ a limit strictly less than 1, hence $\rho^n/E(V^n)$ tends to zero, and so does $\log(1 + (\rho^n/E(V^n)))$.

LEMMA 3. *For some $0 < \rho < e^A$, where $A = \sup(\mathbf{Y} \circ \boldsymbol{\theta} - b(\boldsymbol{\theta}))$, a vector $\boldsymbol{\theta}_0$, and some positive λ_1 and λ_2, the following holds wherever* $\exp(\mathbf{Y} \circ \boldsymbol{\theta} - b(\boldsymbol{\theta})) > \rho$:

$$A - \lambda_1||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 < (\mathbf{Y} \circ \boldsymbol{\theta} - b(\boldsymbol{\theta})) < A - \lambda_2||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2\,.$$

As is well known, the matrix of second-order derivatives of $b(\boldsymbol{\theta})$ is the covariance matrix of y, and hence positive definite. Therefore $\mathbf{Y} \circ \boldsymbol{\theta} - b(\boldsymbol{\theta})$ is strictly convex, and is easily seen to attain its maximum. Let $\boldsymbol{\theta}_0$ be the point where the maximum A is attained. The Taylor expansion of $\mathbf{Y} \circ \boldsymbol{\theta} - b(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$ now yields the stated inequalities for some neighborhood of $\boldsymbol{\theta}_0$, if $2\lambda_1$ and $2\lambda_2$ are larger and smaller than all the eigenvalues of the matrix of second order derivatives of $b(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$. By strict convexity it is now easy to determine $\rho < e^A$ so that it will bound $\exp(\mathbf{Y} \circ \boldsymbol{\theta} - b(\boldsymbol{\theta}))$ outside that neighborhood.

The proposition is now proved by combining the lemmas, and the assumption of local boundedness of the density function of μ_j on $m_j \cap \Theta$.

Qualitatively both our procedure and Akaike's give "a mathematical formulation of the principle of parsimony in model building." Quantitatively, since our procedure differs from Akaike's only in that the dimension is multiplied by $\frac{1}{2}\log n$, our procedure leans more than Akaike's towards lower-dimensional models (when there are 8 or more observations). For large numbers of observations the procedures differ markedly from each other. If the assumptions we made in Section 2 are accepted, Akaike's criterion cannot be asymptotically optimal. This would contradict any proof of its optimality, but no such proof seems to have been published, and the heuristics of Akaike [1] and of Tong [4] do not seem to lead to any such proof.

REFERENCES

[1] AKAIKE, H. (1974). A new look at the statistical identification model. *IEEE Trans. Auto. Control* **19** 716–723.

[2] SCHWARZ, G. (1969). A second order approximation to optimal sampling regions. *Ann. Math. Statist.* **40** 313–315.

[3] SCHWARZ, G. (1971). A sequential Student test. *Ann. Math. Statist.* **42** 1003–1009.

[4] TONG, H. (1975). Determination of the order of a Markov chain by Akaike's information criterion. *J. Appl. Prob.* **12** 488–497.

DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
JERUSALEM
ISRAEL