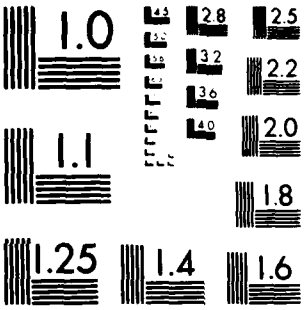
AD-A128 840 USE OF MODEL-SEGMENTATION CRITERIA IN CLUSTERING AND 1/1
SEGMENTATION OF TIME..(U) ILLINOIS UNIV AT CHICAGO
CIRCLE DEPT OF QUANTITATIVE METHODS S L SCLOVE
UNCLASSIFIED 05 MAY 83 UIC/DQM-A83-3 ARO-19085.4-MA F/G 12/1 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

ARO 19085.4-MA
(12)

USE OF MODEL-SEGMENTATION CRITERIA IN CLUSTERING AND SEGMENTATION OF TIME SERIES AND DIGITAL IMAGES

by

STANLEY L. SCLOVE

AD-A120840

To be presented at the
44th Session of the International Statistical Institute,
Madrid, Spain,
September 12-22, 1983

TECHNICAL REPORT NO. UIC/DQM/A83-3
May 5, 1983

PREPARED FOR THE
ARMY RESEARCH OFFICE
UNDER
CONTRACT DAAG29-82-K-0155

Statistical Models and Methods for
Cluster Analysis and Image Segmentation

Principal Investigator: Stanley L. Sclove

DTIC FILE COPY

Reproduction in whole or in part is permitted
for any purpose of the United States Government.
Approved for public release; distribution unlimited

QUANTITATIVE METHODS DEPARTMENT
COLLEGE OF BUSINESS ADMINISTRATION
UNIVERSITY OF ILLINOIS AT CHICAGO
BOX 4348, CHICAGO, IL 60680

DTIC
1983

5/5/83

83 06 01 027

THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

USE OF MODEL-SEGMENTATION CRITERIA IN CLUSTERING AND SEGMENTATION OF TIME SERIES AND DIGITAL IMAGES

STANLEY L. SCLOVE

Department of Quantitative Methods
College of Business Administration
University of Illinois at Chicago

CONTENTS

Abstract

Introduction

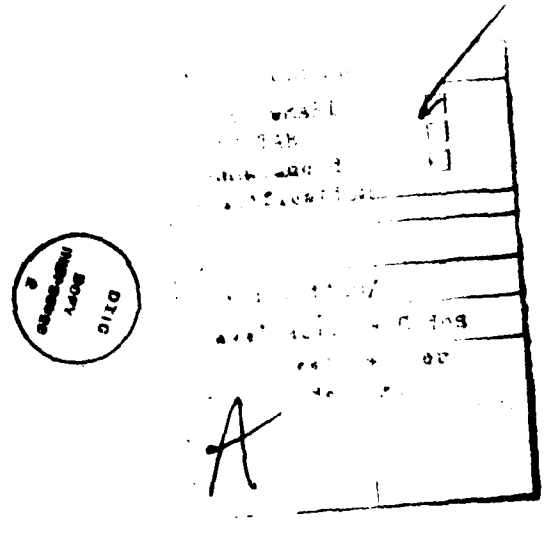
Model-Selection Criteria

Application to Clustering and Segmentation

- Multi-sample clustering
- Mixture-model clustering of individuals
- Segmentation of time series
- Segmentation of digital images

Bibliography

Summary in French



USE OF MODEL-SEGMENTATION CRITERIA IN CLUSTERING AND SEGMENTATION OF TIME SERIES AND DIGITAL IMAGES

STANLEY L. SCLOVE
Department of Quantitative Methods
College of Business Administration
University of Illinois at Chicago

ABSTRACT

This paper treats the development and use of criteria for model selection, particularly for the choice of the number of groups ("clusters") in the analysis of multivariate data and of the number of classes of segments in the segmentation of time series and digital images. Criteria such as those of Akaike, Schwarz and Kashyap are considered.

Key words and phrases: cluster analysis, time-series segmentation, digital image segmentation, model-selection criteria, information criteria, Akaike's information criterion (AIC), Schwarz's criterion

USE OF MODEL-SELECTION CRITERIA IN CLUSTERING AND SEGMENTATION OF TIME SERIES AND DIGITAL IMAGES

S. L. SCLOVE

University of Illinois, Chicago, IL 60680, U.S.A.

INTRODUCTION

This article treats the development and use of model-selection criteria, particularly for the choice of the number of clusters in multivariate data analysis and the number of classes of segment in the segmentation of time series and digital images. Criteria such as those of Akaike (1973, 1974, 1981), Schwarz (1978) and Kashyap (1982) are considered.

MODEL-SELECTION CRITERIA

Consider the problem of choosing from among a number of models, indexed by k (k = 1,2,...,K). Let L(k) be the likelihood given the k-th model. Various model-selection criteria taking the form

$$-2\,\ell n[\max L(k)] + a(n)m(k) + b(k), \tag{1}$$

have been developed in relatively recent years. Here n is the sample size, ℓn denotes the natural logarithm, max L(k) denotes the maximum of the likelihood over the parameters, and m(k) is the number of independent parameters in the k-th model. For a given criterion a(n) is the cost of fitting an additional parameter and b(k) is an additional term depending upon the criterion and the model k.

Akaike, in a very important sequence of papers, including Akaike (1973, 1974, 1981), developed such a criterion as an (heuristic) estimate of the expected entropy (Kullback-Leibler information). Akaike's information criterion (AIC) is of the form (1) with

$$a(n) = 2 \text{ for all } n,\ b(k) = 0 \quad \text{(AIC)}. \tag{2}$$

Schwarz (1978), working from a Bayesian viewpoint, obtained a criterion of the form (1) with

$$a(n) = \ell n\, n,\ b(k) = 0 \quad \text{(Schwarz's criterion)}. \tag{3}$$

Since, for n greater than 8, $\ell n\, n$ exceeds 2, Schwarz's criterion favors models with fewer parameters than does Akaike's. Rissanen (1978a) obtained a criterion of the form (1) as a solution to a problem of minimum-bit representation of a signal. His criterion, for this reason referred to as SDD (shortest data description), is given by

$$a(n) = \ell n[(n+2)/24],\ b(k) = 2\,\ell n(k+1) \quad \text{(Rissanen's criterion)}. \tag{4}$$

Boekee and Buss (1981) studied the performance of several criteria, namely Rissanen's and the criteria given by

$$a(n) = \ell n[(n+2)/24], \quad b(k) = 0 \tag{5}$$

and

$$a(n) = \ell n(n+2), \quad b(k) = 0. \tag{6}$$

Note that (6) is essentially Schwarz's criterion. They simulated a second-order autoregression with autoregression coefficients -0.8 and -0.9 for n=50, 100, 200 and 400 (fifty times for each case) and found that (6), the criterion which is essentially Schwarz's, gave good results, better than the AIC criterion. (It should be mentioned that the b(k) in (4) is specific to the problem of fitting a k-th order autoregression.) The criteria (4) and (5) gave mediocre results, similar to AIC. This assessment by Boekee and Buss was based on the distribution of the order estimate in the simulation experiments (the true value being 2, for second order), for the various criteria.

Note that, of the criteria, only AIC has a(n) a constant function of n. Various researchers, including Kashyap (1982), Rissanen (1978a,b) and Schwarz (1978) have mentioned that AIC is not consistent; a(n) needs to depend upon n. Thus the particular form of (1) chosen by Akaike through his heuristic estimation argument may not be best.

Kashyap (1982), also taking the Bayesian approach, took the asymptotic expansion of the logarithm of the posterior probabilities a term further than did Schwarz and obtained the criterion given by

$$a(n) = \ell n\ n, \quad b(k) = \ell n[\det B(k)] \quad \text{(Kashyap's criterion)}, \tag{7}$$

where det denotes determinant and B(k) is the negative of the matrix of second partials of ℓn L(k), evaluated at the maximum likelihood estimates. In Gaussian linear models this is the covariance matrix of the maximum likelihood estimates of the regression coefficients; in general, the expectation of B(k), evaluated at the true parameter values, is Fisher's information matrix. Since Kashyap's criterion is based on reasoning similar to Schwarz's, but contains an extra term, it could be expected to perform better.

In what follows application of Akaike's, Schwarz's and Kashyap's criteria to various specific problems will be discussed. In these applications often the criteria agree (give the same choice of model), but in cases when they disagree, AIC chooses the least parsimonious model, Kashyap's criterion the most parsimonious, Schwarz's falling in between. Some of the examples studied are of known structure (the correct model is known) and in cases of disagreement Kashyap's criterion did best and Schwarz's second best. Thus, the particular specifications put on the form (1) by Akaike may not be the best. Nonetheless, the profession is greatly in his debt for repeatedly calling our attention to the very important model-selection problem.

APPLICATION TO CLUSTERING AND SEGMENTATION

Multi-sample clustering

The problem of multi-sample clustering, the grouping of samples, is treated in Bozdogan and Sclove (1982), where numerical examples are given. The situation is the K-sample problem (one-way analysis of variance), with an emphasis on grouping the samples into fewer than K clusters. The use of model-

selection criteria in this situation can provide an alternative to multiple-comparison procedures. Use of model-selection criteria avoids the difficult choice of levels of significance in such problems. Here in the Gaussian case with p variables one has a mean vector for each population. With separate covariance matrices, $m(k) = k[p + p(p+1)/2]$. With a common covariance matrix $m(k) = kp + p(p+1)/2$. Model-selection criteria can also be used to decide whether or not to assume a common covariance matrix.

Mixture-model clustering of individuals

Bozdogan (1983) applies model-selection criteria to the choice of the number of populations in the population mixture model. (See, e.g., Wolfe 1970.) Here there are k-1 independent mixture probabilities. In the Gaussian case with p variables and different covariance matrices, $M(k) = k-1 + k[p + p(p+1)/2]$. The algorithm and computer programs of Wolfe (1970) can be used to obtain the maximum likelihood estimates for fixed k. Then model-selection criteria can be used to estimate k.

Segmentation of Time Series

A model for clustering or segmentation is given by assuming that each instance of observation gives rise not only to an observation x but also to a label γ, equal to 1, 2, ..., or k, where k is the number of classes. Model-selection criteria are used to estimate k. In the context of this model, clustering is merely estimation of the labels. Sclove (1982a,c) treats the problem of segmentation of time series by modeling the label process as a Markov chain. An algorithm and computer programs are discussed; numerical examples are given. The parameters are the transition probabilities, the marginal probabilities of the classes, and the parameters of the class-conditional densities, so m(k) can be taken to be $k(k-1) + (k-1) + c(k)$, where $c(k) = k[p + p(p+1)/2]$ in the Gaussian case with separate covariance matrices.

Segmentation of Digital Images

Similar ideas are applied to digital images in Sclove (1982a,b). Here the label process is modeled as a one-sided Markov random field, In the first-order case the label of each pixel is conditioned on the labels immediately to the north and west of it. The number of independent transition probabilities is $k^2(k-1)$. Further details and examples are given in Sclove (1982a).

BIBLIOGRAPHY

Akaike, H.(1973). Information theory and an extension of the maximum likelihood principle. *Proc. 2nd International Symposium on Information Theory*, pp. 267-281. Akademia Kiado, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control 6*, pp. 716-723.

Akaike, H. (1981). Likelihood of a model and informa-tion criteria. *J. Econometrics 16*, pp. 3-14.

Boekee, D.E., and BUSS, H.H. (1981). Order estimation of autoregressive models. *4th Aachener Kolloquium: Theorie und Anwendung der Signalverarbei-*

Trans, pp. 126-130.

Bozdogan, H.(1983). Determining the number of component clusters in standard multivariate normal mixture model using model-selection criteria. To appear as Technical Report UIC/DQM/A83-1, Army Research Office Contract DAAG29-82-K-0155, S.L. Sclove, Principal Investigator, University of Illinois at Chicago.

Bozdogan, H., and Sclove, S.L.(1982). Multi-sample cluster analysis using Akaike's information criterion. Technical Report A82-2, Army Research Office Contract DAAG29-82-K-0155, University of Illinois at Chicago (submitted to *Annals of Statistical Mathematics*).

Kashyap, R.L.(1982). Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Trans. Pattern Analysis and Machine Intelligence 4*, pp. 99-104.

Rissanen, J.(1978a). Modelling by shortest data description. *Automatica 14*, pp. 465-471.

Rissanen, J.(1978b). Consistent order estimates of autoregressive processes by shortest description of data. *International Symposium on Optimization and Analysis of Stochastic Systems*, Oxford, pp. 1-11.

Schwarz, G.(1978). Estimating the dimension of a model. *Annals of Statistics 6*, pp. 461-464.

Sclove, S.L.(1982a). On segmentation of time series and images in the signal detection and remote sensing contexts. Technical Report 82-4, Office of Naval Research Contract N00014-80-C-0408, S.L. Sclove, Principal Investigator, University of Illinois at Chicago. To appear in *Proc. Workshop on Signal Processing in the Ocean Environment*, Marcel-Dekker, Inc., 1983.

Sclove, S.L.(1982b). Application of the conditional population-mixture model to image segmentation. Technical Report A82-1, Army Research Office Contract DAAG29-82-K-0155, University of Illinois at Chicago. To appear in *IEEE Trans. Pattern Analysis and Machine Intelligence*.

Sclove, S.L.(1982c). Time-series segmentation: a model and a method. Technical Report A82-3, Army Research Office Contract DAAG29-82-K-0155, University of Illinois at Chicago. To appear in *Information Sciences*.

Wolfe, J.H.(1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research 5*, pp. 329-350.

RESUMÉ

L'UTILISATION DES CRITÈRES POUR LA SÉLECTION DES MODÈLES DANS LAS RÉPARTITION ET LA SEGMENTATION DES SÉRIES TEMPORELLES ET DES IMAGES NUMÉRICALES

Cet article traite le développement et l'utilisation des critères pour la sélection des modèles, surtout pour le choix du nombre des groupes (c'est-à-dire des "clusters") dans l'analyse des données multidimensionnelles et du nombre des classes des segments dans la segmentation des séries temporelles et des images numériqales. Les critères comme ceux de Akaike, Schwarz et Kashyap y sont considérés.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1 REPORT NUMBER: Technical Report No. UIC/DQM/A83-3	2. GOVT ACCESSION NO.	3 RECIPIENT'S CATALOG NUMBER
4 TITLE (and Subtitle): Use of Model-Selection Criteria in Clustering and Segmentation of Time Series and Digital Images		5 TYPE OF REPORT & PERIOD COVERED: Technical Report 6 PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Stanley L. Sclove		8 CONTRACT OR GRANT NUMBER(s): DAAG29-82-K-0155
9. PERFORMING ORGANIZATION NAME AND ADDRESS: University of Illinois at Chicago Box 4348, Chicago, IL 60680		10 PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE: May 5, 1983 13 NUMBER OF PAGES: 4 + ii
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15 SECURITY CLASS. (of this report): Unclassified 15a DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17 DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

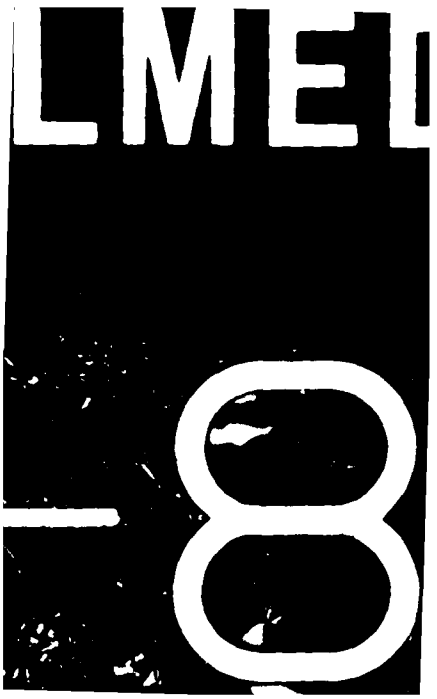
cluster analysis, time-series segmentation, digital image segmentation, model-selection criteria, information criteria, Akaike's information criterion (AIC), Schwarz's criterion

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper treats the development and use of criteria for model selection, particularly for the choice of the number of groups ("clusters") in the analysis of multivariate data and of the number of classes of segments in the segmentation of time series and digital images. Criteria such as those of Akaike, Schwarz and Kashyap are considered.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)


LMEL
8