PSYCHOMETRIKA—VOL. 52, NO. 3, 333–343
SEPTEMBER 1987
SPECIAL SECTION

APPLICATION OF MODEL-SELECTION CRITERIA TO SOME PROBLEMS IN MULTIVARIATE ANALYSIS

STANLEY L. SCLOVE
UNIVERSITY OF ILLINOIS AT CHICAGO

A review of model-selection criteria is presented, with a view toward showing their similarities. It is suggested that some problems treated by sequences of hypothesis tests may be more expeditiously treated by the application of model-selection criteria. Consideration is given to application of model-selection criteria to some problems of multivariate analysis, especially the clustering of variables, factor analysis and, more generally, describing a complex of variables.

Key words: model selection, model evaluation, Akaike's information criterion, AIC, Schwarz's criterion, cluster analysis, clustering variables, factor analysis.

Model Selection Versus Hypothesis Testing

Model-selection criteria take account simultaneously of both the goodness-of-fit (likelihood) of a model and the number of parameters used to achieve that fit. The criteria take the form of a penalized likelihood function, that is, the negative log likelihood plus a penalty term, which increases with the number of parameters.

Often one is in an exploratory phase of data analysis, trying to find a model. For example, one may be planning to compare two treatment means but allowing possibly unequal population variances. Then, rather than doing a preliminary test of equality of variances, one could consider and evaluate four different models.

Multiple comparison of means is another situation in which thinking in terms of evaluation of the members of a set of alternative models may be more appropriate than a sequence of hypothesis tests. The possible clusterings constitute alternative models. For example, if there are three groups, then the possible clusterings are as in Table 2.

Fitting a polynomial is another such situation. In the simplest case, that of fitting a quadratic or a linear function, instead of testing the hypothesis of nullity of the coefficient of the quadratic term, why not assign figures of merit to the linear and quadratic models and choose between them accordingly? There is an obvious generalization here to fitting higher-order polynomials and to choosing the order of an autoregressive scheme.

Test procedures involve choice of significance levels; application of test procedures to multiple-decision problems involves choice of a number of dependent significance levels. Also, as Akaike states (1987), test procedures do not penalize for overparametrization because usually a saturated model is used as a reference.

Hypothesis testing is well suited to some problems, where Type I and Type II error probabilities have frequency interpretations in actual populations, for example, treatment populations in a clinical setting, or when in sampling inspection for quality assurance the error rates correspond to consumers' and producer's risk. But, often model-selection will provide a viable alternative to hypothesis testing.

Requests for reprints should be sent to Stanley L. Sclove, Department of Information and Decision Sciences, College of Business Administration, University of Illinois at Chicago, Box 4348, Chicago, IL 60680-4348.

0033-3123/87/0900-SS02$00.75/0

TABLE 1

Four Different Models for Comparison of Means, Allowing Inequality of Variances

Model	Means	Variances	Number of parameters
1	equal	equal	2
2	equal	unequal	3
3	unequal	equal	3
4	unequal	unequal	4

Review of Model-Selection Criteria

This section of the paper treats some aspects of the development and application of criteria for model-selection. Criteria such as those of Akaike (1973, 1974, 1981, 1983), Schwarz (1978) and Kashyap (1982) are considered.

Consider, then, the problem of choosing from among a number of models, indexed by k ($k = 1, 2, \ldots. K$), and let $L(k)$ be the likelihood (joint probability density function of the observations, considered as a function of the parameters, with the observations held fixed) given the k-th model. Various model-selection criteria taking the form

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

have been developed in relatively recent years. Here n is the sample size, "log" 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

TABLE 2

The Five Possible Clusterings of Three Objects with Application to Multiple Comparison of Three Means

Model	Clustering	Number of parameters (means)
1	{1,2,3}	1
2	{1,2}, {3}	2
3	{1,3}, {2}	2
4	{2,3}, {1}	2
5	{1}, {2}, {3}	3

$a(n)$ is the cost of fitting an additional parameter and $b(k, n)$ is an additional term depending upon the criterion and the model k. A model is good if it gives a small value of (1), relative to the values given by competing models.

Akaike, in an important sequence of papers, including Akaike (1973, 1974, 1981, 1983), developed such a criterion as an estimate of the expected entropy (Kullback-Leibler information). Akaike's information criterion (AIC) is of the form (1) with

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

Schwarz (1978) and Kashyap (1982) work from the Bayesian viewpoint, expanding Pr (Model $k \mid$ data), the posterior probability of model k, given the data. The multiplier of $m(k)$ in this expansion is log n; that is, in Schwarz's and Kashyap's criteria, $a(n) = \log n$, rather than $a(n) = 2$ as in AIC. So Schwarz's criterion is of the form (1) with

$$a(n) = \log n, \quad b(k, n) = 0. \tag{3}$$

Since, for n greater than 8, log n exceeds 2, Schwarz's criterion favors models with fewer parameters than does Akaike's. Kashyap's criterion will be discussed more fully shortly. First let us note other developments yielding criteria similar to (3). Rissanen (1978) 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

$$a(n) = \log\left[\frac{n+2}{24}\right], \qquad b(k, n) = 2\log(k+1) \quad \text{(Rissanen's criterion).} \tag{4}$$

See also Rissanen (1983, 1985).

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

$$a(n) = \log\left[\frac{n+2}{24}\right], \qquad b(k, n) = 0, \tag{5}$$

and

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

Note that (6) is essentially Schwarz's criterion. Boekee and Buss 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). Table 3 summarizes their results; some of the contents of Table 3 appear in their paper, but here several columns, those for the means, mean squared errors and square root of mean squared errors, have been appended. The results indicate that (6), the criterion which is essentially Schwarz's, gave good results, better than the AIC criterion. (It should be mentioned that the $b(k, n)$ 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 (1978, 1980) and Schwarz (1978) have mentioned that AIC is not consistent; for consistency $a(n)$ needs to depend upon n. However, consistency is an asymptotic property, and any real problem has a finite sample size n.

Since there are different choices of $a(n)$, suppose for example that one considered the

TABLE 3

Simulation Results of Boekee and Buss (1981), Comparing Four Model-Selection Criteria

	Estimate of order:								
	1	2	3	4	5				
	Frequencies:					Total	Mean	MSE	Sq.root of MSE
n= 50									
AIC	0	33	10	6	1	50	2.500	.860	.927
1A	0	44	5	1	0	50	2.140	.180	.424
1B	0	14	10	13	4	41	3.171	2.390	1.546
1C	0	20	12	11	3	46	2.935	1.804	1.343
n=100									
AIC	0	35	10	3	2	50	2.440	.800	.894
1A	0	43	7	0	0	50	2.140	.140	.374
1B	0	20	16	10	4	50	2.960	1.840	1.356
1C	0	36	9	3	2	50	2.420	.780	.883
n=200									
AIC	0	40	9	1	0	50	2.220	.260	.510
1A	0	47	3	0	0	50	2.060	.060	.245
1B	0	40	9	1	0	50	2.220	.260	.510
1C	0	40	9	1	0	50	2.220	.260	.510
n=400									
AIC	0	39	8	3	0	50	2.280	.400	.632
1A	0	50	0	0	0	50	2.000	.000	.000
1B	0	47	3	0	0	50	2.060	.060	.245
1B	0	49	1	0	0	50	2.020	.020	.141

1A: a(n) = log(n+2), b(k,n) = 0
1B: a(n) = log[(n+2)/24], b(k,n) = 0
1C: a(n) = log[(n+2)/24], b(k,n) = 2 log(k+1)
AIC: a(n) = 2, b(k,n) = 0

criterion given by $a(n) = 2$ and that given by $a(n) = 3$ to be reasonable. These choices of $a(n)$ give the criteria

$$g(k; 2) = -2 \log [\max L(k)] + 2m(k),$$

and

$$g(k; 3) = -2 \log [\max L(k)] + 3m(k).$$

The choice $a(n) = 2$, that is, the criterion $g(k; 2)$, favors larger models, since it costs only 2 units, rather than 3, to fit an extra parameter. If a small model is best according to $g(k; 2)$, then it will be best also according to $g(k; 3)$; if a large model is best according to $g(k; 3)$, it will be best according to $g(k; 2)$. Thus, by use of two criteria, a range of reasonable values of $m(k)$ results. The minimum plausible number of parameters, in the situation where just these two criteria are considered, is that number favored by $g(k; 3)$; the maximum plausible number of parameters is that favored by $g(k; 2)$.

More generally, consider alternative model-selection criteria taking the form

$$g(k) = -2 \log [\max L(k)] + a(n)m(k). \tag{7}$$

Different approaches lead to different choices of $a(n)$; however, for consistency the fastest increasing is $a(n) = \log n$ and the slowest increasing is $a(n) = \log (\log n)$. Define

$$l(n) = \min \{2, \log \log n, \log n\} = \min \{2, \log \log n\},$$

and

$$u(n) = \max \{2, \log \log n, \log n\}, = \max \{2, \log n\}.$$

Use of $a(n) = l(n)$ favors large models; use of $a(n) = u(n)$ favors small models. If a small model is best according to $l(n)$, it is globally best, because it will also be best according to $u(n)$. If a large model is best according to $u(n)$, it is globally best, because it will also be best according to $l(n)$.

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

$$a(n) = \log n, \qquad b(k, n) = \log [\det \mathbb{B}(k, n)], \tag{8}$$

where det denotes determinant and $\mathbf{B}(k, n)$ is the negative of the matrix of second partials of $\log L(k)$, 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 $\mathbf{B}(k, n)$, evaluated at the true parameter values, is Fisher's information matrix. Kashyap's derivation is perhaps somewhat more general than Schwarz's, in that Schwarz's development is explicit only for exponential families. Since Kashyap's criterion is based on reasoning similar to Schwarz's, but contains an extra term, it could be expected to perform better.

Next, a review of some applications of model-selection criteria to various specific problems will be given.

Multisample Clustering

The procedure of multisample clustering, the grouping of samples, is treated in Bozdogan (1981, 1986), and Bozdogan and Sclove (1984); numerical examples and discussion are given in these papers. Stated briefly, 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 and 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 can be used to obtain the maximum-likelihood estimates for fixed k. Then model-selection criteria can be used to estimate k. In this context some analyticity conditions required for series expansions yielding the model-selection criteria are not met, and though the criteria can be regarded as heuristic figures of merit, more research is required.

Segmentation of Time Series

A model for clustering or segmentation of time series is given by assuming that each instance of observation gives rise not only to an observation x but also to a label, 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 (1983b, 1984) 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 $g(k - 1) + (k - 1) + c(k)$, 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 (1983a, 1984). In Sclove (1984) 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 (1984). Note the large number of parameters in this model. A review of Akaike's development of AIC shows that he is to a great extent motivated by such large-scale problems, for example, multivariate time series.

Clustering Variables

Heuristic algorithms have been applied to the problem of clustering variables, based on use of correlations as similarity measures and applying the maximum, minimum, or average pairwise correlation as a measure of similarity between two clusters of variables. Also, variables can be clustered by using principal components of factors as axes and plotting the variables, using their loadings as coordinates. Here an alternative approach is suggested. Why not consider various configurations of the covariance matrix as corresponding to clusterings of variables, and evaluate the special configurations as alternative models, using model-selection criteria?

The approach is to identify each clustering with an appropriately restricted covariance (or correlation) matrix.

Let us consider in detail the special cases $p = 2$ and $p = 3$ variables. In the case of two variables the possible clusterings are Model 1: $\{1, 2\}$, and Model 2: $\{1\}, \{2\}$. Model 1 has an unrestricted covariance matrix but Model 2 has Cov(1, 2) = 0. For multivariate normal samples, model-selection criteria of the form (7) become

$$g(k) = np \log(2\pi) + n \log [\det \mathbf{S}_k] + np + a(n)m(k), \tag{9}$$

where $\mathbf{S}_k$ is the maximum-likelihood estimate of the covariance matrix under model k. This is just the matrix of sample covariances (with divisors of n), except that a zero replaces the sample covariance in the restricted positions.

The restriction imposed by Model 2 is that the variables be uncorrelated, that is, their covariance is set equal to zero rather than being estimated by its maximum likelihood estimate. One chooses Model 1 over Model 2 if $g(1)$ is less than $g(2)$. The difference

$g(1) - g(2)$ can be written as

$$n \log \left(\frac{\det \mathbf{S}_1}{\det \mathbf{S}_2}\right) + a(n),$$

which equals

$$n \log(1 - r^2) + a(n),$$

where r is the sample correlation, so that Model 1 is better than Model 2 if $r^2 > 1 - \exp(-a(n)/n)$. For example, with $a(n) = 2$ (as in AIC) and $n = 8$, this is $r^2 > 1 - \exp(-1/4)$, or about .22, or absolute value of $r > .47$. Such a value, if viewed as a cut-off point for a hypothesis test, implies a significance level. A value $r^2 = .22$ with $n = 8$ corresponds to a t-value of 0.55 with 6 d.f.; the corresponding one-tailed P-value is between .25 and .40. When $a(n) = \log n$ (Schwarz's criterion), this becomes $r^2 > 1 - n^{-1/n}$. If $a(n)$ were proportional to n, say $a(n) = cn$, then this would become $r^2 > 1 - \exp(-c)$, a constant not dependent upon n.

For three variables the covariance matrix contains 6 parameters. The clustering $\{1, 2, 3\}$ (all three variables in a single cluster) uses all six parameters, placing no restrictions on them. The clustering $\{1\}, \{2\}, \{3\}$ (each variable forming its own cluster) corresponds to restricting the covariance matrix to be diagonal and hence uses only 3 of 6 possible parameters, the off-diagonal elements (covariances) in the dispersion matrix being restricted to zero. The clustering $\{1, 2\}, \{3\}$ corresponds to Cov(1, 2) nonzero, Cov(1, 3) = 0 and Cov(2, 3) = 0 and hence involves only 4 rather than 6 free parameters. Similarly for the clusterings $\{1, 3\}, \{2\}$ and $\{1\}, \{2, 3\}$.

One could score all $2^3 = 8$ models resulting from setting each of Cov(1, 2), Cov(1, 3) and Cov(2, 3) equal to zero or not. Five of these eight models correspond to the above clusterings. The other three do not correspond to clusterings. For example, setting only Cov(2, 3) equal to zero means only that Variables 2 and 3 should not be in the same cluster and could be viewed as no clustering or as being consistent with all three clusterings $\{1, 2\}, \{3\}$; $\{1, 3\}, \{2\}$; and $\{1\}, \{2\}, \{3\}$.

More generally, for p variables, one approach is to score all $2^{p(p-1)/2}$ models resulting from setting each of the $p(p - 1)/2$ covariances equal to zero or not. Some of these will indicate clusterings, others will not.

An important feature of the application of model-selection criteria to clustering is that clustering is not forced; the alternative of no clustering is evaluated along with all the others. In the present context this means that $\{1, 2, 3\}$, representing no clustering (i.e., all variables in the same cluster) is evaluated along with the other possibilities. This deals with a frequent criticism of clustering algorithms, namely, that they force clustering on the data.

Setting one or another of the off-diagonal elements equal to zero may affect the requirement that the covariance matrix be positive definite (p.d.). A simple example of this is given by the case of a correlation matrix for $p = 3$ variables with equal correlations, r. The condition for positive definiteness is $r > -1/2$. Now suppose the (1, 3)-element is changed to zero. The resulting matrix is p.d. iff (i) $1 - r^2 > 0$, and (ii) the determinant is positive, that is, $1 - 2r^2 > 0$. It is still given that $r > -\frac{1}{2}$, so Condition (i) is satisfied, but (ii) requires $-\frac{1}{2}^{1/2} < r < \frac{1}{2}^{1/2}$. We have $r > -\frac{1}{2}^{1/2}$ (because it is still given that $r > -\frac{1}{2}$), but $r < \frac{1}{2}^{1/2}$ is a further restriction. If the (2, 3)-element as well as the (1, 3)-element is set equal to zero, it produces a covariance matrix corresponding to the clustering $\{1, 2\}, \{3\}$, and this matrix is p.d., regardless of the value of r.

In theory, then, the approach given would be improved by imposing some conditions to insure that the covariance matrix will remain p.d. when setting one or another of the off-diagonal elements equal to zero. The need for conditions to insure positive definiteness could limit somewhat the applicability of the technique; only those clusterings corresponding to a set of p.d. matrices are allowable. (It may indeed be the case that those restrictions corresponding to actual clusterings do maintain positive definiteness; this point needs further research.) In practice there is no problem (except for computer time) with trying all possible patterns of off-diagonal zeros, letting the computer program first check the current pattern for positive definiteness (e.g., check that all eigenvalues are positive). If not, skip the current pattern of zeros and go on to the next one.

The notion of clustering variables treated here is that separate clusters imply zero covariances. Within clusters, some of the correlations may be negative. This admits easy interpretation when the negatives are removable by reflection; in other cases the interpretation may be problematic.

Example

Consider clustering $p = 3$ variables; systolic blood pressure, diastolic blood pressure, and height. The correlations are given near the beginning of Table 4. These correlations are consistent with those between these three variables for the first 20 cases in Table 2-2a of Dixon and Massey (1969). The results of clustering of these variables are given in Table 4. This example was chosen for illustration because it is small, so that a list of results for all possible cases can be shown. Also, it may be helpful to show a situation where one feels he can guess the answer; here it seems clear that the two blood pressure measurements should be in one cluster and height in the other, since (presumably) height is not strongly related to blood pressure (at least not in the general population; there may be particular syndromes characteristic of people at one or the other extreme of height where the blood pressure is abnormally high or low). Table 4 gives the correlations and the values of AIC and Schwarz's criterion. The clustering {Sys, Dias}, {Ht} is indeed that chosen by both criteria.

Factor Analysis

A model for factor analysis can be written

$$\mathbf{x} = \Lambda\mathbf{f} + \mathbf{u},$$

where $\mathbf{x}$ is the p-variate observation, $\mathbf{f}$ is the vector of m uncorrelated factors ($m < p$), taken to have unit variance, and $\mathbf{u}$ is the vector of residuals, the variances of which are called the uniquenesses. The variables in $\mathbf{u}$ are uncorrelated with one another and with the variables in $\mathbf{f}$. The p-by-m column-orthogonal matrix Λ consists of the factor loadings.

An important problem is the choice of the number of factors, m.

According to this model, the covariance matrix Σ of x takes the form

$$\Sigma = \Lambda\Lambda' + \Psi,$$

where Ψ, the covariance matrix of $\mathbf{u}$, is diagonal, its diagonal elements being the uniquenesses. Thus factor analysis can be viewed as fitting restricted estimates of Σ, of this specified form. The number of parameters is $mp + p$, since Ψ has p elements and there are m factors, each having p loadings. The condition that Λ be column-orthogonal imposes $m(m - 1)/2$ constraints. Thus the number of free parameters is $mp + p - m(m - 1)/2$. In order for a factor analysis to make sense, the number m of factors must be limited so that this quantity is less than $p(p + 1)/2$, the number of variances and covariances. Equating these two expressions gives a logical upper limit, say M, for m. The choice of the number of factors as being 1, 2, ..., or M is a multiple-decision problem—a choice between alternative models—to which model-selection criteria can reasonably be applied. This matter is discussed in Akaike (1987) and Bozdogan and Ramirez (1987).

TABLE 4
Results for Example of Clustering of Three Variables

Clustering of the p = 3 Variables,
Systolic and Diastolic Blood Pressure and Height

Sample covariance matrix, S:

354.800	164.179	4.158
164.179	127.713	4.039
4.158	4.039	6.513

Sample correlation matrix, R:

1.000	.771	.086
.771	1.000	.140
.086	.140	1.000

					Model-Selection Criterion:		
I1	I2	I3	Minus 2 Log Likelihood	No. of Parameters	AIC	Schwarz Criterion	Clustering
0	0	0	422.176	3	428.176	431.163	{Sys},{Dias},{Ht}
0	0	1	421.780	4	429.780	433.763	{Sys},{Dias,Ht}
0	1	0	422.026	4	430.026	434.009	{Sys,Ht},{Dias}
0	1	1	422.011	5	432.011	436.989	---
1	0	0	404.105	4	412.105	416.088	{Sys,Dias},{Ht}
					-------	-------	---------------
1	0	1	404.082	5	414.082	419.061	---
1	1	0	404.672	5	414.672	419.651	---
1	1	1	403.686	6	415.686	421.660	{Sys,Dias,Ht}

```
COV(1,2) = 0       IF I1 = 0
         = S(1,2)  IF I1 = 1
COV(1,3) = 0       IF I2 = 0
         = S(1,3)  IF I2 = 1
COV(2,3) = 0       IF I3 = 0
         = S(2,3)  IF I3 = 1
```

Describing a Complex of Variables

The goal in describing a complex of variables could be stated as attempting to arrive at a maximally parsimonious description or summarization of the set of variables. This could involve factor analysis, or clustering variables, or both.

Given the above view of clustering variables, each clustering corresponds to a restricted estimate of Σ. Factor analysis also corresponds to fitting restricted estimates of Σ. This means that factor analysis and clustering variables can be simultaneously scored by model-selection criteria. To do this, note that, given a sample from a p-variate normal distribution, one has

$$-2 \log \max L(k) = np \log(2\pi) + n \log \det \mathbf{S}_k + n \operatorname{tr} (\mathbf{S}_k^{-1}\mathbf{S}),$$

where tr denotes the trace of a matrix, $\mathbf{S}$ is the sample covariance matrix (the unrestricted maximum-likelihood estimate of Σ), and the subscript k ranges over the union of the set of factor analysis models and the set of models corresponding to the possible clusterings of variables.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *2nd International Symposium on Information Theory* (pp. 267–281). Budapest: Akademia Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 6,* 716–723.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics, 16,* 3–14.

Akaike, H. (1983). Statistical inference and measurement of entropy. In H. Akaike & C.-F. Wu (Eds.), Scientific inference, data analysis, and robustness (pp. 165–189). New York: Academic Press.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52.*

Boekee, D. E., & Buss, H. H. (1981). Order estimation of autoregressive models. *4th Aachener Kolloquium: Theorie und Anwendung der Signalverarbeitung* [Proceedings of the 4th Aachen Colloquium: Theory and application of signal processing]. (pp. 126–130).

Bozdogan, H. (1981). Multi-sample cluster analysis and approaches to validity studies in clustering individuals. Unpublished doctoral dissertation, University of Illinois at Chicago, Department of Mathematics, Chicago.

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

Bozdogan, H. (1986). Multi-sample cluster analysis as an alternative to multiple comparison procedures. *Bulletin of Informatics and Cybernetics, 22* (No. 1-2), 95–130.

Bozdogan, H., & Ramirez, D. E. (1987). An expert model selection approach to determine the "best" pattern structure in factor analysis models. Unpublished manuscript.

Bozdogan, H., & Sclove, S. L. (1984). Multi-sample cluster analysis using Akaike's information criterion. *Annals of Institute Statistical Mathematics, 36,* 163–180.

Dixon, W. J., & Massey, F. J. (1969). Introduction to statistical analysis (3rd ed.). New York: McGraw-Hill.

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

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14,* 465–471.

Rissanen, J. (1980). Consistent order estimates of autoregressive processes by shortest description of data. In O. L. R. Jacobs, M. H. A. Davis, M. A. H. Dempster, C. J. Harris, & P. C. Parks (Eds.), *Analysis and Optimisation of Stochastic Systems* (pp. 451–461). London and New York: Academic Press.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics, 11,* 416–431.

Rissanen, J. (1985). Minimum-description-length principle. *Encyclopedia of Statistical Sciences* (Vol. 5, pp. 523–527). New York: John Wiley & Sons.

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

Sclove, S. L. (1983a). Application of the conditional population-mixture model to image segmentation. *IEEE Transactions Pattern Analysis and Machine Intelligence, 5,* 428–433.

Sclove, S. L. (1983b). Time-series segmentation: A model and a method. *Information Sciences, 29,* 7–25.

Sclove, S. L. (1984). On segmentation of time series and images in the signal detection and remote sensing contexts. In E. W. Wegman & J. G. Smith (Eds.), *Statistical signal processing* (pp. 421–434). New York: Marcel Dekker.

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