PSYCHOMETRIKA—VOL. 52, NO. 3, 317–332
SEPTEMBER 1987
SPECIAL SECTION

FACTOR ANALYSIS AND AIC

HIROTUGU AKAIKE

THE INSTITUTE OF STATISTICAL MATHEMATICS

The information criterion AIC was introduced to extend the method of maximum likelihood to the multimodel situation. It was obtained by relating the successful experience of the order determination of an autoregressive model to the determination of the number of factors in the maximum likelihood factor analysis. The use of the AIC criterion in the factor analysis is particularly interesting when it is viewed as the choice of a Bayesian model. This observation shows that the area of application of AIC can be much wider than the conventional i.i.d. type models on which the original derivation of the criterion was based. The observation of the Bayesian structure of the factor analysis model leads us to the handling of the problem of improper solution by introducing a natural prior distribution of factor loadings.

Key words: factor analysis, maximum likelihood, information criterion AIC, improper solution, Bayesian modeling.

1. Introduction

The factor analysis model has been producing thought provoking statistical problems. The model is typically represented by

$$\mathbf{y}(n) = A\mathbf{x}(n) + \mathbf{u}(n), \qquad n = 1, 2, \cdots, N$$

where $\mathbf{y}(n)$ denotes a p-dimensional vector of observations, $\mathbf{x}(n)$ a k-dimensional vector of factor scores, $\mathbf{u}(n)$ a p-dimensional vector of specific variations. It is assumed that the variables with different n's are mutually independent and that $\mathbf{x}(n)$ and $\mathbf{u}(n)$ are mutually independently distributed as Gaussian random variables with variance covariance matrices $I_{k \times k}$ and Ψ, respectively, where Ψ is a diagonal matrix. The covariance matrix Σ of $\mathbf{y}(n)$ is then given by

$$\Sigma = AA' + \Psi.$$

This model is characterized by the use of a large number of unknown parameters, much larger than the number of unknown parameters of a model used in the conventional multivariate analysis. The empirical principle of parsimony in statistical model building dictates that the increase of the number of parameters should be stopped as soon as it is observed that a further increase does not produce significant improvement of fit of the model to the data. Thus the control of the number of parameters has usually been realized by applying a test of significance.

The author would like to express his thanks to Jim Ramsay, Yoshio Takane, Donald Ramirez and Hamparsum Bozdogan for helpful comments on the original version of the paper. Thanks are also due to Emiko Arahata for her help in computing.

Requests for reprints should be sent to Hirotugu Akaike, The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-Ku, Tokyo 106, Japan.

0033-3123/87/0900-SS01$00.75/0

In the case of the maximum likelihood factor analysis this is done by adopting the likelihood ratio test. However, in this test procedure, the unstructured saturated model is always used as the reference and the significance is judged by referring to a chi-square distribution with a large number of degrees of freedom equal to the difference between the number of parameters of the saturated model and that of the model being tested. As will be seen in section 3, an example discussed by Jöreskog (1978) shows that direct application of such a test to the selection of a factor analysis model is not quite appropriate. There the expert's view clearly contradicts the conventional use of the likelihood ratio test.

In 1969 the present author introduced final prediction error (FPE) criterion for the choice of the order of an autoregressive model of a time series (Akaike, 1969, 1970). The criterion was defined by an estimate of the expected mean square one-step ahead prediction error by the model with parameters estimated by the method of least squares. The successful experience of application of the FPE criterion to real data suggested the possibility of developing a similar criterion for the choice of the number of factors in the factor analysis. The choice of the order of an autoregression controlled the number of unknown parameters in the model, that controlled the expected mean square one-step ahead prediction error. By analogy it was easily observed that the control of the number of factors was required for the control of the expected prediction error by the fitted model. However, it was not easy to identify what the prediction error meant in the case of the factor analysis.

In the case of the autoregressive model an estimate of the expected predictive performance was adopted as the criterion; in the case of the maximum likelihood factor analysis it was the fitted distribution that was evaluated by the likelihood. The realization of this fact quickly led to the understanding that our prediction was represented by the fitted model in the case of the factor analysis, which then led to the understanding that the expectation of the log likelihood with respect to the "true" distribution was related to the Kullback-Leibler information that defined the amount of deviation of the "true" distribution from the assumed model.

The analogy with the FPE criterion then led to the introduction of the criterion

$$\text{AIC} = (-2) \log \text{maximum likelihood} + 2 \text{(number of parameters)},$$

as the measure of the badness of fit of a model defined with parameters estimated by the method of maximum likelihood, where log denotes a natural logarithm (Akaike, 1973, 1974). We will present a simple explanation of AIC in the next section and illustrate its use by applying it to an example in section 3.

Although AIC produces a satisfactory solution to the problem of the choice of the number of factors, the application of AIC is hampered by the frequent appearance of improper solutions. This shows that successive increase of the number of factors quickly lead to models that are not quite appropriate for the direct application of the method of maximum likelihood.

In section 4 it will be discussed that the factor analysis model may be viewed as a Bayesian model and the choice of a factor analysis model by minimizing the AIC criterion is essentially concerned with the choice of a Bayesian model. This recognition encourages the use of further Bayesian modeling for the elimination of improper solutions. In section 5 a natural prior distribution for the factor loadings is introduced through the analysis of the likelihood function. Numerical examples will be given in Section 6 to show that the introduction of the prior distribution suppresses the appearance of improper solutions and that the indefinite increase of a communality caused by the conventional maximum likelihood procedure may be viewed as of little practical significance.

The paper concludes with brief remarks on the contribution of factor analysis to the development of general statistical methodology.

2. Brief Review of AIC

The fundamental ideas underlying the introduction of AIC are:

1. The predictive use of the fitted model.
2. The adoption of the expected log likelihood as the basic criterion.

Here the concept of parameter estimation is replaced by the estimation of a distribution and the accuracy is measured by a universal criterion, the expected log likelihood of the fitted model.

The relation between the expected log likelihood and the Kullback-Leibler information number is given by

$$I(f;\, g) = E \log f(x) - E \log g(x),$$

where $I(f;\, g)$ denotes the Kullback-Leibler information of the distribution f relative to the distribution g, and E denotes the expectation with respect to the "true" distribution $f(x)$ of x. The second term on the right-hand side represents the expected log likelihood of an assumed model $g(x)$ with respect to the "true" distribution $f(x)$. Since $I(f;\, g)$ provides a measure of the deviation of f from g and since $\log g(x)$ provides an unbiased estimate of $E \log g(x)$ the above equation provides a justification for the use of log likelihoods for the purpose of comparison of statistical models.

Consider the situation where the model $g(x)$ contains unknown parameter θ, that is, $g(x) = g(x \mid \theta)$. When the data x are observed and the maximum likelihood estimate $\theta(x)$ of θ is obtained, the predictive point of view suggests the evaluation of $\theta(x)$ by the goodness of $g(\cdot \mid \theta(x))$ as an estimate of the true distribution $f(\cdot)$. By adopting the information $I(f;\, g)$ as the basic criterion we are led to the use of $E_y \log g(y \mid \theta(x))$ as the measure of the goodness of $\theta(x)$, where E_y denotes the expectation with respect to the true distribution $f(y)$ of y. To relate this criterion to the familiar log likelihood ratio test statistic we adopt $2E_y \log g(y \mid \theta)$ as our measure of the goodness of $g(y \mid \theta)$ as an estimate of $f(y)$.

Here we consider the conventional setting where the true distribution $f(y)$ is given by $g(y \mid \theta_0)$, that is, θ_0 is the true value of the unknown parameter, the data x are a realization of the vector of i.i.d. random variables $x_1, x_2, \cdots, x_N$, and the log likelihood ratio test statistic asymptotically satisfies the relation

$$2 \log g(x \mid \theta(x)) - 2 \log g(x \mid \theta_0) = \chi^2_m,$$

where χ^2_m denotes a chi-squared with degrees of freedom m which is equal to the dimension of the parameter vector θ. Under this setting it is expected that the curvature of the log likelihood surface provides a good approximation to that of the expected log likelihood surface. This observation leads to another asymptotic equality

$$2E_y \log g(y \mid \theta(x)) - 2E_y \log g(y \mid \theta_0) = -\chi^2_m,$$

where it is assumed that y is another independent observation from the same distribution as that of x and the chi-squared variable is identical to that defined by the log likelihood ratio test statistic.

The above equations show that the amount of increase of $2 \log g(x \mid \theta(x))$ from $2 \log g(x \mid \theta_0)$ obtained by adjusting the parameter value by the method of maximum likelihood is asymptotically equal to the amount of decrease of $2E_y \log g(y \mid \theta(x))$ from $2E_y$ log

$g(y \mid \theta_0)$. Thus, to measure the deviation of $\theta(x)$ from θ_0 in terms of the basic criterion of twice the expected log likelihood, χ^2_m must be subtracted twice from 2 log $g(x \mid \theta(x))$ to make the difference of twice the log likelihoods an unbiased estimate of that of twice the expected log likelihoods.

Since χ^2_m is unobservable, as we do not know θ_0, we consider the use of its expected value m. The negative of the quantity thus obtained defines

$$\mathrm{AIC} = (-2) \log g(x \mid \theta(x)) + 2m.$$

When several different g's are compared the one that gives the minimum of AIC represents the best fit. Such an estimate is denoted as MAICE (minimum AIC estimate). For more detailed discussion of the predictive point of view of statistics and the use of the information criterion readers are referred to Akaike (1985).

3. How AIC Works With The Factor Analysis Model

Given a set of observations $y = (y(n); n = 1, 2, \cdots, N)$ the maximum likelihood factor analysis starts with the definition of the log likelihood function given by

$$\log L(k) = -\tfrac{1}{2}N[\log |\Sigma_k| + \mathrm{tr}\, \Sigma_k^{-1}S],$$

where S denotes the sample covariance matrix of y and k the number of factors and Σ_k is given by

$$\Sigma_k = A_k A_k' + \Psi,$$

where A_k denotes the matrix of factor loadings and Ψ the uniqueness variance matrix. The diagonal elements of $A_k A_k'$ define the communalities. The AIC statistic for the k-factor model is then defined by

$$\mathrm{AIC}(k) = (-2) \log L(k) + [2p(k + 1) - k(k - 1)].$$

To show the use of AIC in the maximum likelihood factor analysis and to illustrate the difference between the AIC and conventional test approach in particular here we will discuss an example treated by Jöreskog (1978, p.457). This examle is concerned with the analysis of Harman's example of twenty-four psychological variables. The unrestricted four factor model was first fitted which produced

$$\chi^2_{186} = 246.36.$$

This model was considered to be "representing a reasonably good fit" but a further restriction of parameters produced a simple structure model with

$$\chi^2_{231} = 301.42.$$

This model was accepted as the best fitting simple structure.

Now we have

$$\mathrm{Prob}\ \{\chi^2_{186} \geqq 246.36 \mid H_0\} \approx 0.0009,$$

and

$$\mathrm{Prob}\{\chi^2_{231} \geqq 301.42 \mid H_0'\} \approx 0.0005,$$

where H_0 and H_0' denote the hypotheses of the four factor and the simple structure, respectively, and the chi-squared variables stand for the random variables with respective degrees of freedom. By the standard of conventional tests these figures show that the results are extremely significant and both H_0 and H_0' should be rejected. In spite of this, the expert judgment of Jöreskog was to accept the four factor model as a reasonable fit and prefer the simple structure model to the unrestricted. This conclusion suggests that the large values of the degrees of freedom appearing in the chi-squared statistics preclude the application of conventional levels of significance, such as 0.05 or 0.01, in making the final judgment of models in this situation.

The chi-squared statistic is defined by

$$\chi^2 = (-2) \max \log L(H) - (-2) \max \log L(H_\infty),$$

where max log $L(H)$ denotes the maximum log likelihood under the hypothesis H and H_∞ denotes the saturated or completely unconstrained model. Since AIC for an hypothesis H is defined by

$$\mathrm{AIC}(H) = (-2) \max \log L(H) + 2 \dim \theta,$$

where dim θ denotes the dimension of the vector of unknown parameters θ, we have

$$\mathrm{AIC}(H) - \mathrm{AIC}(H_\infty) = \chi^2_{\mathrm{d.f.}} - 2(\mathrm{d.f.}),$$

where d.f. denotes the difference between the number of unknown parameters of H_∞ and that of H. By neglecting the common additive constant AIC (H_∞) we may define AIC(H) simply by

$$\mathrm{AIC}(H) = \chi^2_{\mathrm{d.f.}} - 2(\mathrm{d.f.}).$$

For the models discussed by Jöreskog we get

$$\begin{aligned}\mathrm{AIC}(H_0) &= 246.36 - 2 \times 186\\ &= -125.64,\end{aligned}$$

and

$$\begin{aligned}\mathrm{AIC}(H_0') &= 301.42 - 2 \times 231\\ &= -160.58.\end{aligned}$$

Since AIC(H_∞) = 0, these AIC's show that both H_0 and H_0' are by far better than H_∞ and that the simple structure model H_0' is showing a better fit than the unrestricted four factor model H_0.

This result by AIC is in complete agreement with Jöreskog's conclusion. The conventional theory of statistics does not tell how to evaluate the significance of a test in each particular application and there is no hope of arriving at a similar conclusion. Obviously the objective procedure of model selection by an information criterion can be fully implemented to define an automatic factor analysis procedure. Such a possibility is discussed by Bozdogan and Ramirez (1987).

4. Factor Analysis Model Viewed as a Bayesian Model

As was demonstrated by the application to Jöreskog's example the AIC approach produced a satisfactory solution to the model selection problem in factor analysis. In spite of this success the use of AIC in the maximum likelihood factor analysis has been severely limited by the frequent occurrence of improper solutions, that is, by the appearance of zero estimates of specific variances. Apparently this is caused by the overparametrization of the model.

The introduction of AIC is motivated by the desire to control the effect of overparametrization and the minimum AIC procedure for model selection is considered to be a realization of the well-known empirical principle of parsimony in statistical modeling. However the application of the minimum AIC procedure assumes the existence of proper maximum likelihood estimates of the models considered. The frequent occurrence of improper solutions in the maximum likelihood factor analysis means that the models are often too much overparametrized for the application of the method of maximum likelihood. This suggests the necessity of further control of the likelihood function. This can be realized by the use of some proper Bayesian modeling.

Before going into the discussion of this Bayesian modeling we will first notice the essentially Bayesian characteristic of the factor analysis model and point out that the minimum AIC procedure is concerned with the problem of the selection of a Bayesian model. In the basic factor analysis model $\mathbf{y} = A\mathbf{x} + \mathbf{u}$ the vector of observations $\mathbf{y}$ is assumed to be distributed following a Gaussian distribution with mean $A\mathbf{x}$ and unique variance Ψ. The vector of factor scores x is unobservable but is assumed to be distributed following a Gaussian distribution with zero mean and variance $I_{k \times k}$. Since $\mathbf{x}$ is never observed this distribution is simply a psychological construction for the explanation of the behavior of y. Under the assumption that A is fixed the distribution of $\mathbf{x}$ specifies the prior distribution of the mean of the observation $\mathbf{y}$. Thus we can see that the choice of k, the number of factors, is essentially concerned with the choice of a Bayesian model. Incidentally, the recognition of the Bayesian characteristic of the factor analysis model also suggests the use of the posterior distribution of $\mathbf{x}$ for the estimation of the factor scores as is discussed by Bartholomew (1981).

The basic problem in the use of a Bayesian model is how to justify the use of a subjectively constructed model. Our belief is that it is possible only by considering various possibilities as alternative models and comparing them with an objectively defined criterion. In particular we propose the use of the log likelihood, or the AIC when some parameters are estimated by the method of maximum likelihood, as the criterion of fit.

Let us consider the likelihood of a factor analysis model as a Bayesian model. For a Bayesian model specified by the data distribution $p(\cdot \mid \theta)$ and prior distribution $p(\theta)$ its likelihood with respect to the observed data y is given by

$$\int p(y \mid \theta) p(\theta) \, d\theta.$$

From the representation $\mathbf{y}(n) = A\mathbf{x}(n) + \mathbf{u}(n)$, $n = 1, 2, \cdots, N$, and the assumption of the mutual independence among the variables the likelihood of the Bayesian model defined with $\theta = (x(1), x(2), \cdots, x(N))$ is given by

$$L = \prod_{n=1}^{N} \left(\frac{1}{2\pi}\right)^{p/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2} \operatorname{tr} \Sigma^{-1} \mathbf{y}(n)\mathbf{y}(n)'\right\}$$

$$= \left(\frac{1}{2\pi}\right)^{Np/2} |\Sigma|^{-N/2} \exp\left\{-\frac{N}{2} \operatorname{tr} \Sigma^{-1} S\right\},$$

where $\Sigma = AA' + \Psi$, $|\Sigma|$ denotes the determinant, and

$$S = \frac{1}{N} \sum_{n=1}^{N} \mathbf{y}(n)\mathbf{y}(n)'.$$

For simplicity the mean of $\mathbf{y}(n)$ is assumed to be zero. Thus we get

$$\log L = -\frac{N}{2} [\log |\Sigma| + \operatorname{tr} S\Sigma^{-1}] + \text{const.}$$

This is exactly the likelihood function used in the conventional maximum likelihood factor analysis. Thus the maximum likelihood estimates of A and Ψ in the classical sense are the maximum likelihood estimates of the unknown parameters of a Bayesian model.

The above result shows that the AIC criterion defined for the factor analysis model is actually the ABIC criterion for the evaluation of a Bayesian model with parameters estimated by the method of maximum likelihood, where ABIC is defined by (Akaike, 1980)

$$\begin{aligned}\text{ABIC} = (-2)&\text{ maximum log likelihood of a Bayesian model}\\&+ 2\text{ (number of estimated parameters).}\end{aligned}$$

In the case of the factor analysis model we have

$$\text{ABIC} = \text{AIC}.$$

This identity clearly shows that there is no essential distinction between the classical and Bayesian models when they are viewed from the point of view of the information criterion.

5. Control of Improper Solutions by a Bayesian Modeling

The appearance of improper solutions suggests the necessity of the reduction of the number of parameters to be estimated by the method of maximum likelihood. The recognition of the Bayesian structure of the factor analysis model suggests that further modeling of the prior distribution of the unknown parameters in A and Ψ is possible. The use of the Bayesian approach for the control of improper solutions is already discussed in an earlier paper by Martin and McDonald (1975). These authors point out the importance of choosing a prior distribution that does not have the appearance of arbitrariness and discuss the use of a reasonably defined prior distribution of specific variances.

The informational approach to statistics puts very much faith in the information supplied by the log likelihood. Hence in the present paper we try to develop a prior distribution without using outside information except for the knowledge of the likelihood function of the data distribution. In the present situation this is particularly appropriate as the prior distribution is considered only for the purpose of tempering the likelihood function to clarify the nature of improper solutions.

By this approach we need a detailed analysis of the likelihood function. For the convenience of the analysis let us consider

$$q = \left(-\frac{2}{N}\right) \log L - \log |S|,$$

where the log likelihood $\log L$ is defined in the preceding section. By ignoring the additive constant we have

$$q = -\log |\Sigma^{-1}S| + \operatorname{tr} \Sigma^{-1}S.$$

By putting $\Psi = D^2$, where D is a diagonal matrix with positive diagonal elements, we get

$$\begin{aligned}\Sigma &= AA' + D^2\\ &= D(I + CC')D,\end{aligned}$$

where A is $p \times k$, I is a $p \times p$ identity matrix and $C = D^{-1}A$, the matrix of standardized factor loadings. We have

$$\mathrm{tr}\ \Sigma^{-1}S = \mathrm{tr}\ (I + CC')^{-1}D^{-1}SD^{-1},$$

and

$$|\Sigma^{-1}S| = |(I + CC')^{-1}||D^{-1}SD^{-1}|.$$

The modified negative log likelihood q can conveniently be expressed by using the eigenvectors z_i and eigenvalues ζ_i of $D^{-1}SD^{-1}$, the standardized sample covariance matrix. Define the matrix Z by

$$Z = [z_1, z_2, \cdots, z_p].$$

It is assumed that Z is normalized so that $Z'Z' = I$ holds. Represent C by Z in the form

$$C = ZF.$$

Adopt the representation

$$FF' = \sum_{i=1}^{p} \mu_i m_i m_i',$$

where $\mu_i > 0$, for $i = 1, 2, \cdots, k$, $= 0$, otherwise, and $m_i' m_i = \delta_{ij}$, where $\delta_{ij} = 1$, for $i = j$, 0, otherwise. Then we get

$$CC' = ZFF'Z' = \sum_{i=1}^{p} \mu_i l_i l_i',$$

where $l_i = Zm_i$ with $l_i' l_j = \delta_{ij}$, and

$$I + CC' = \sum_{i=1}^{p} \lambda_i l_i l_i',$$

where $\lambda_i = 1 + \mu_i$. From this representation we get

$$(I + CC')^{-1} = \sum_{i=1}^{p} \lambda_i^{-1} l_i l_i',$$

and

$$\mathrm{tr}\ (I + CC')^{-1}D^{-1}SD^{-1} = \mathrm{tr} \sum_i \lambda_i^{-1} l_i l_i' \sum_j \zeta_j z_j z_j'$$
$$= \sum_i \sum_j \lambda_i^{-1} \zeta_j m_i^2(j),$$

where $m_i(j)$ denotes the j-th element of m_i. The last relation is obtained from the equation

$$z_j' l_i = m_i(j).$$

We also have

$$|(I + CC')^{-1}|\,|D^{-1}SD^{-1}| = \prod_{i=1}^{p} \lambda_i^{-1} \prod_{j=1}^{p} \zeta_j.$$

Thus we get the following representation of the modified negative likelihood function as a function of $\lambda = (\lambda_{1,} \lambda_2, \cdots, \lambda_p)$ and $m = (m_1, m_2, \cdots, m_p)$:

$$q(\lambda, m) = -\sum_{i=1}^{p} \log \lambda_i^{-1} \zeta_i + \sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_i^{-1} \zeta_j m_i^2(j).$$

Assume that ζ_i and λ_i are arranged in the descending order, that is, $\zeta_1 \geqq \zeta_2 \geqq \cdots \geqq \zeta_p$ and $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_p$, where $\lambda_{k+1} = \cdots = \lambda_p = 1$. Then the successive minimization of $q(\lambda, m)$ with respect to $m_p, m_{p-1}, \cdots, m_1$ leads to

$$q(\lambda) = \sum_{i=1}^{k} [\lambda_i^{-1}\zeta_i - \log (\lambda_i^{-1}\zeta_i)] + \sum_{i=k+1}^{p} (\zeta_i - \log \zeta_i).$$

As a function of λ, $(\lambda^{-1}\zeta) - \log (\lambda^{-1}\zeta)$ attains its minimum at $\lambda = \zeta$, for $\zeta > 1$, and at $\lambda = 1$, otherwise. Thus we get

$$\underset{\lambda}{\text{Min}}\ q(\lambda) = k^* + \sum_{i=k^*+1}^{p} (z_i - \log \zeta_i),$$

where $\zeta_i > 1$, for $i \leqq k^*$, $\leqq 1$, otherwise. This last quantity is equal to the quantity given by the Equation (18) of Jöreskog (1967, p.448) and is $(-2/N)$ times the maximum log likelihood of the factor analysis model when D is given.

In maximizing the likelihood we would normally hope that a too small value of some of the diagonal elements of D will reduce the maximum likelihood of the corresponding model. However, that this is not the case is shown by the above result which explains that the value of the maximum likelihood is sensitive only to the behavior of smaller eigenvalues of $D^{-1}SD^{-1}$. A very small diagonal element of D will only produce a very large eigenvalue. Thus the process of maximizing the likelihood with respect to the elements of D does not eliminate the possibility of some of these elements going down to zero.

The form of $q(\lambda)$ shows that if we introduce an additive term $\rho\Sigma\mu_i$ with $\rho > 0$ then the minimization of

$$q(\lambda) = \sum_{i=1}^{p} [\lambda_i^{-1}\zeta_i - \log (\lambda_i^{-1}\zeta_i)] + \rho \sum_{i=1}^{p} \mu_i,$$

with respect to λ does not allow any of λ_i $(= 1 + \mu_i)$ going to infinity. Taking into account the relations $C = ZF$ and $FF' = \Sigma\mu_i m_i m_i'$ we get

$$\sum_{i=1}^{p} \mu_i = \text{tr } FF' = \text{tr } CC'.$$

Since $C = D^{-1}A$ the minimization of $q(\lambda)$ produces an estimate that is given as the posterior mode under the assumption of the prior distribution given by

$$K \exp \left\{ -\frac{N}{2} \rho \text{ tr } D^{-1}AA'D^{-1} \right\},$$

where K denotes the normalizing constant and N the sample size. This prior distribution is defined by a spherical normal distribution of the standardized factor loadings and will be referred to as the standard spherical prior distribution of the factor loadings.

For the complete specification of the Bayesian model it is necessary to define the prior distribution of D. However, an arbitrarily defined prior distribution of the elements of D can easily eliminate improper solutions if only it penalizes smaller values sufficiently. Since our interest here is mainly in the clarification of the nature of improper solutions obtained by the conventional maximum likelihood procedure we will not proceed to the modeling of the prior distribution of D and simply adopt the uniform prior.

TABLE 1

Communality estimates*

Harman : eight physical

$\rho = 0$ (MLE)

$k \backslash i$	1	2	3	4	5	6	7	8
1	842	865	810	813	240	171	123	199
2	830	893	834	801	911	636	584	463
3	872	1000	806	844	909	641	589	509

$\rho = 0.1$

$k \backslash i$	1	2	3	4	5	6	7	8
1	837	858	804	810	241	172	124	200
2	828	881	828	800	855	647	591	476
3	858	910	830	832	859	650	590	523
4	865	910	832	843	851	689	649	521
5				same as above				

$\rho = 1.0$

$k \backslash i$	1	2	3	4	5	6	7	8
1	763	768	725	739	252	181	134	204
2	766	781	742	743	590	486	440	409
3								
4				same as above				
5								

* In this and following tables maximum possible communality is normalized to 1000.

6. Numerical Examples

The Bayesian model defined with the standard spherical prior distribution of the factor loadings was applied to six published examples of improper solutions. These examples are Harman's eight physical variables data (Harman, 1960, p.82), with $p = 8$ and improper at $k = 3$, Davis data (Rao, 1955, p.110), with $p = 9$ and improper at $k = 2$, Maxwell's normal children data (Maxwell, 1961, p.55), with $p = 10$ and improper at $k = 4$, Emmett data (Lawley & Maxwell, 1971, p.43), with $p = 9$ and improper at $k = 5$, Maxwell's neurotic children data (p.53), with $p = 10$ and improper at $k = 5$, and Harman's twenty-four psychological variables data (Harman, 1960, p.137), with $p = 24$ and improper at $k = 6$.

The informational point of view suggests that hyperparameter ρ of the prior distri-

TABLE 2

Communality estimates

Davis data

$\rho = 0$ (MLE)

$k \backslash i$	1	2	3	4	5	6	7	8	9
1	658	661	228	168	454	800	705	434	703
2	652	1000	243	168	464	816	704	435	701
3	1000	661	220	204	451	1000	701	488	696

$\rho = 0.1$

$k \backslash i$	1	2	3	4	5	6	7	8	9
1	653	656	226	167	451	790	700	431	697
2	694	689	227	171	470	800	698	434	696
3	701	695	251	197	470	801	698	444	696
4					same as above				

$\rho = 1.0$

$k \backslash i$	1	2	3	4	5	6	7	8	9
1	596	598	210	156	415	702	633	400	631
2									
3					same as above				
4									

bution may be "estimated" by maximizing the likelihood of the Bayesian model with respect to ρ. However, for this purpose integration in a high-dimensional space is required. In this paper we will limit our attention to the analysis of solutions with some fixed values of ρ.

The estimation of specific variances under the present Bayesian model was realized by the following procedure. Given the initial estimate D_1^2 of D^2 the sample covariance matrix S is replaced by $S_1 = D_1^{-1}SD_1^{-1}$ and the next estimate D_2^2 of D^2 is obtained by the relation $D_2^2 = \text{diag}(S - D_1B_1B_1'D_1)$, where B_1 is a $p \times k$ matrix such that B_1B_1' provides a least squares fit

$$2\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\Big[S_1(i,j) - \sum_{l=1}^{k}B_1(i,l)B_1(j,l)\Big]^2 + \rho\sum_{i=1}^{p}\sum_{j=1}^{k}B_1^2(i,j) = \text{Min.},$$

where $B(i,j)$ denotes (i,j)th element of B. The estimates of communalities are defined by diag $(D_1B_1B_1'D_1)$. The process is repeated until convergence is established.

When $\rho = 0$ the above procedure produced maximum likelihood solutions that were confirmed by a procedure based on the result of Jennrich and Robinson (1969). When $\rho > 0$ the solution may only be considered as an arbitrary approximation to the posterior

TABLE 3

Three factor maximum likelihood solution of Emmett data

i	A(•1)	A(•2)	A(•3)	Ψ
1	.664	.321	.074	.450
2	.689	.247	-.193	.427
3	.493	.302	-.222	.617
4	.837	-.292	-.035	.212
5	.705	-.315	-.153	.381
6	.819	-.377	.105	.177
7	.611	.396	-.078	.400
8	.458	.296	.491*	.462
9	.766	.427	-.012	.231

* The value suggests singular increase of the 8th communality.

mode. Nevertheless it will be sufficient for the purpose of confirming of the effect of the tempering of the likelihood function. For convenience we will call the solution the Bayesian estimate.

In the case of the above six examples the choice of $\rho = 1.0$ produced solutions with signficant overall reduction of communalities, or increase of specific variances. With the choice of $\rho = 0.1$ solutions were usually close to the conventional maximum likelihood estimates but with the improper estimates of communalities suppressed. Improper estimates disappeared completely, unless ρ was made extremely small. For a fixed ρ estimates of communalities usually stabilized as k, the number of factors, was increased.

It was generally observed that when the maximum likelihood method produced an improper solution first at $k = k_0$ the corresponding Bayesian estimate with $\rho = 0.1$ was proper but with only one communality estimate inflated compared with the estimate at $k = k_0 - 1$. Such a singular increase of the communality means the reinterpretation of a part of the specific variation as an independent factor. This fact and the result of our analysis of the likelihood function suggest that the singular increase of the communality is usually caused by the overparametrization that makes the estimate sensitive to the sampling variability of the data rather than by the structural change of the best fitting model at $k = k_0$. This is in agreement with the earlier observation of Tsumura and Sato (1981) on the nature of improper solutions.

Tables 1 and 2 provide estimates of communalities of Harman's eight physical variables data and of Davis' data, respectively, for various choices of the order, k, and ρ. In the

TABLE 4

Communality estimates by various procedures

Emmett data

$\rho = 0$ (MLE)

$k\backslash i$	1	2	3	4	5	6	7	8	9
1	510	537	300	548	390	481	525	224	665
2	538	536	332	809	592	778	597	256	782
3	550	573	384	788	619	823	600	538	769
4	554	666	379	772	663	856	648	480	759
5	556	868	1000	780	664	836	666	464	743

$\rho = 0.1$

$k\backslash i$	1	2	3	4	5	6	7	8	9
1	502	529	296	545	388	478	516	221	652
2	535	531	330	790	588	762	590	252	753
3	549	561	378	783	611	786	590	399	750
4	same as above								
5									

$\rho = 1.0$

$k\backslash i$	1	2	3	4	5	6	7	8	9
1	425	448	254	478	344	422	434	189	540
2	433	450	261	522	391	472	445	196	551
3	same as above								
4									
5									

case of the Harman data the result in Table 1 shows that the improper value 1000 at $i = 2$ with $k = 3$, obtained with $\rho = 0$, disappeared for the positive values of ρ. In particular, with $\rho = 0.1$, the solutions with $k = 2$, 3 and 4 are all mutually very close and they are close to the solutions with $\rho = 0$ and $k = 2$ and 3, except for the improper component at $k = 3$. This suggests that the two-factor model is an appropriate choice, which is in agreement with Harman's original observation. The soltuion with $\rho = 1.0$ conforms with this observation.

For the Davis data with $k_0 = 2$ the non-uniqueness of the convergence of iterative procedures for the maximum likelihood was first reported by Tsumura, Fukutomi, and Asoo (1968). With $k = 2$, Jöreskog (1967, p.474) reported improper estimate of specific variance for the 1st component and Tsumura et al. (p.57) found one for the 8th component. As is shown in Table 2 our procedure found one at the 2nd component. The result

TABLE 5

Suggested choices of dimensionalities*

Harman:	eight physical	
$p=8$	$k_o=3$	$k_s=2$
	MAICE $=\infty$**	
Davis		
$p=9$	$k_o=2$	$k_s=1$
	MAICE $=\infty$**	
Maxwell:	normal	
$p=10$	$k_o=4$	$k_s=3$
	MAICE $=\infty$**	
Emmett		
$p=9$	$k_o=5$	$k_s=2$
	MAICE $=3$	
Maxwell:	neurotic	
$p=10$	$k_o=5$	$k_s=2$
	MAICE $=3$	
Harman :	24 variables	
$p=24$	$k_o=6$	$k_s=5$
	MAICE $=5$	

* p : dimension of observation
k_o : lowest order with improper solution
k_s : suggested order by the Bayesian analysis

** ∞ denotes saturated model.

given in Table 4 of Martin and McDonald (1975, p.515) also suggests the existence of improper solution with zero unique variance for the 2nd component. These results suggest the existence of local maxima of the likelihood function. Table 2 also gives improper estimates for the 1st and 6th components with $k = 3$, which is in agreement with the result reported by Jöreskog.

The estimates obtained with $\rho = 0.1$ may be viewed as practically identical and are close to the solution with $\rho = 0$, the maximum likelihood estimate, for $k = 1$. This result strongly suggests that the improper solutions are spurious in the sense that they can be suppressed by mild tempering of the likelihood function. The one-factor model seems a reasonable choice in this case. The solution with $\rho = 1.0$ conforms with the present observation.

The phenomenon of the singular increase of a communality estimate is observed even with $k < k_0$. Such an example is given by the three-factor maximum likelihood solution of the Emmett data. The maximum likelihood solution by Lawley and Maxwell (1971, p.43) is reproduced in Table 3 which suggests the singular increase of the communality of the 8th component at $k = 3$. In Table 4 the estimate with $\rho = 0.1$ shows substantial increase of communality at only the 8th component at $k = 3$, compared with the estimate at $k = 2$. The increase is completely suppressed with $\rho = 1.0$. This result suggests that the high value of the communality estimate of the 8th component at $k = 3$ obtained with $\rho = 0$ is spurious. Similar phenomenon was observed with Maxwell's data of neurotic children for the 2nd component at $k = 3$.

Tsumura and Sato (1981, p.163) report that, by their experience, improper solutions were always with "quasi-specific factors" that respectively showed singular contributions to some specific variances. The above example shows that our present Bayesian approach can detect the appearance of such a factor even before one gets a definitely improper solution. Thus we can expect that the present approach will realize a reasonable control of improper solutions.

Table 5 summarizes the suggested choices of the number of factors for the six examples where the choices by the minimum AIC procedure, MAICE, are also included. The suggested choices are based on subjective judgments of the numerical results. It is quite desirable to develop a numerical procedure for the evaluation of the likelihood of each Bayesian model to arrive at an objective judgment.

It is interesting to note here that by a proper choice of ρ the Bayesian approach can produce estimate of A even with $k = p$. This explains the drastic change of the emphasis between the modelings by the conventional and Bayesian approach. By the Bayesian approach there is no particular meaning in trying to reduce the number of factors. To avoid unnecessary distortion of the model it is even advisable to adopt a large value of k and control the estimation procedure by a proper choice of ρ.

7. Concluding Remarks

It is remarkable that the idea of factor analysis has been producing so much stimulus to the development of statistical modeling. In terms of the structure of the model it is essentially Bayesian. Nevertheless, the practical use of the model was realized by the application of the method of maximum likelihood and this eventually led to the introduction of AIC.

The concept of the information measure underlying the introduction of AIC leads our attention from parameters to the distribution. This then provides a conceptual framework for the handling of the Bayesian modeling as a natural extension of the conventional statistical modeling. The occurence of improper solutions in the maximum likelihood factor analysis is a typical example that explains the limitation of the conventional modeling. The introduction of the standard spherical prior distribution of factor loadings provided an example of overcoming the limitation by a proper Bayesian modeling.

This series of experiences clearly explains the close dependence between the factor analysis and AIC, or the informational point of view of statistics, and illustrates their contribution to the development of general statistical methodology. It is hoped that this close contact between psychometrics and statistics will be maintained in the future and contribute to the advancement of both fields.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics, 21,* 243–247.

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics, 22,* 203–217.

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: Akademiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19,* 716–723.

Akaike, H. (1980). Likelihood and the Bayes procedure. In J. M. Bernardo, M. H. De Groot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian Statistics* (pp. 143–166). Valencia: University Press.

Akaike, H. (1985). Prediction and entropy. In A. C. Atkinson & S. E. Fienberg (Eds.), *A Celebration of Statistics* (pp. 1–24). New York: Springer-Verlag.

Bartholomew, D. J. (1981). Posterior analysis of the factor model. *British Journal of Mathematical and Statistical Psychology, 34,* 93–99.

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

Harman, H. H. (1960). *Modern Factor Analysis.* Chicago: University Press.

Jennrich, R. I., & Robinson, S. M. (1969). A Newton-Raphson algorithm for maximum likelihood factor analysis. *Psychometrika, 34,* 111–123.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32,* 443–482.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43,* 443–477.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method, 2nd Edition.* London: Butterworths.

Martin, J. K., & McDonald, R. P. (1975). Bayesian estimation in unrestricted factor analysis: a treatment for Heywood cases. *Psychometrika, 40,* 505–517.

Maxwell, A. E. (1961). Recent trends in factor analysis. *Journal of the Royal Statistical Society, Series A, 124,* 49–59.

Rao, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika, 20,* 93–111.

Tsumura, Y., Fukutomi, K., & Asoo, Y. (1968). On the unique convergence of iterative procedures in factor analysis. *TRU Mathematics, 4,* 52–59. (Science University of Tokyo).

Tsumura, Y., & Sato, M. (1981). On the convergence of iterative procedures in factor analysis. *TRU Mathematics, 17,* 159–168. (Science University of Tokyo).