Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301478

A MODIFIED SCHWARZ CRITERION FOR INCOMPLETE DATA

HASSANIA HAMZAOUI*
Sidi Mohamed Ben Abdellah University, Faculty of Sciences, Fez.
Corresponding Author Email: hassania.hamzaoui@usmba.ac.ma

ABDELAZIZ ALOUI
Sidi Mohamed Ben Abdellah University, Faculty of Sciences, Fez.
Email: abdelaziz.aloui@usmba.ac.ma

ABDELAZIZ EL MATOUAT
Le Havre University, Normandie, Le Havre. Email: abdelaziz.el-matouat@univ-lehavre.fr

Abstract

It is well-known that the selection criteria make it possible to determine the order of a statistical model associated with the observed data. But in practice, the problem of missing values requires a modification of these criteria. For Akaike Information criterion, this problem of incomplete data was studied by Cavanaugh and Shumway (1998), they demonstrated an extension of Akaike's criterion to take account of missing values. But this criterion does not always lead to correct model selection. In this paper, we propose a new information criterion of Schwarz. This criterion is based on the motivation provided for the posterior probability of the candidate model and the EM algorithm. We have validated the theoretical results on simulated data. The new criterion converges to the correct order of the candidate model for both small and large samples, even if the percentage of missing data increases.

Keywords: Incomplete Data, Model Selection, Information Criteria, a Posterior Probability, Em Algorithm.

1. INTRODUCTION

Suppose we have a statistical structure (Ω, P_θ, $\theta \in \Theta$) where P_θ is a probability absolutely continuous about a measurement of Lebesgue and Θ a convex set of unknown dimension k_o, then we have a family of probability densities $f(.|\theta)$ such as for

$\theta \in \Theta$:

$$f(.|\theta) = \frac{dP_\theta}{dx}$$

Let us consider also a sequence of parameterized models M_1, M_2... M_L associated with a sample of data Y. Assuming that each Mk is uniquely parametrized by a vector θ_k, presumed to lie in a parameter space Θ. Our objective is to estimate the dimension k_o of the vector parameter called order of the model when the sample Y contains the missing value. Indeed, according to the nature of the problem to study, the observed data can be incomplete. This situation is met, for example, in the study of the production of a company, or in genetics where the problem of the missing values is due to the autofecondation which is sometimes impossible to observe. The problem of model selection is widely resolved by information criteria.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301478

In the complete data, the general form of these criteria is written:

$$IC(k) = -2lnf\left(Y\middle|\hat{\theta}_n^k\right) + kC_n$$

$$\hat{k} = \underset{1\leq k\leq L}{\operatorname{argmin}} IC(k)$$

n is the sample size of Y, $\hat{\theta}_n^k$ denotes the estimator of θk obtained by maximizing the likelihood and C_n a factor of penalization allowing to attenuate the entropic over−parameterization of the model related to a criterion based on log− likelihood. For C_n = 2, we obtain the criterion of Akaike noted AIC, which is one of the most popular and effective criterion used for model selection but it does not always end in a satisfactory estimate.

It implies a strict over−parametrization of the order [Shibata, [10]]. For C_n = ln (n), we obtain the criterion of Schwarz which provides a consistent estimator of the order. Let us suppose that Y is an incomplete data in its general form; it implies the existence of two sample spaces Y_{obs} the actual values and Y_{mis} the missing part. In selecting models from data Y, Cavanaugh and Shumway [5] used the criterion of Shimodaira noted PDIO (”Predictive Divergence for Incomplete Observation Models”) [11] and EM algorithm to derive a modified criterion of Akaike, noted AIC_{cd} (’cd’ indicate ’complete data’). Although, it takes the incomplete aspect of the data into account, this criterion remains nonconvergent according to the results of simulation presented in the last part of this paper.

We propose to use the EM algorithm and the posterior probability of the candidate model Mk to derive a new criterion of the type SIC noted SIC_{cd} characterized by a significant penalization of the entropy of the missing values and log−likelihood of the actual values, and allowing improvement of the criterion AIC_{cd}.

Finally we compare the criteria AIC_{cd}, AIC, SIC and SIC_{cd} thanks to a study of a simulated causal autoregressive model. We check in particular the new criterion leads to a correct estimation of the order of small ones and large samples as well as that for a significant number of missing data, thus validing the improvement of criterion AIC_{cd}.

2. STUDY OF THE INCOMPLETE DATA *Y*

Now, we suppose that a given incomplete sample Y where the observations are re-ordered, we note Y= (Y_{obs}, Y_{mis}) where Y_{obs} and Y_{mis} are respectively the parts observed and missing from Y.

2.1 The Expectation − Maximization (EM) Algorithm

The EM algorithm was proposed in (1976) by Dempster Laird Rubin [6]. It is an iterative application to estimate the parameters. It optimizes the probability of a statistical model M on the condition of using a class of distributions often associated to the exponential family.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301478

We denote as Q the parametric function to optimize defined by:

$$Q(\theta|\theta') = \int lnf(Y|(M,\theta))f(Y_{mis}|Y_{obs},(M,\theta'))dY_{mis}$$

Where f ($Y_{mis}|Y_{obs}$, (M, θ′)) denotes the parametric density function conditioned to an observed data. The EM algorithm is reiterated in two steps: initially, we calculate the density f ($Y_{mis}|Y_{obs}$, (M, θ′)) and after, we update the parameter estimated by computing a standard Maximum Likelihood according to the observed data.

These two steps represent only one iteration. At the end of each iteration, the θ′ optimal ones are substituted from the θ by indexing the θ and θ′ according to the various iterations; we obtain the mechanism $\widehat{\theta'}_k \rightarrow \theta_{k+1}$.

The function Q guarantees growth of the function probability and the checking of the conditions of regularity, and consequently the function:

$$V_n(\theta^k) = -\frac{1}{n}\ln f\Big(Y\Big|(M_k,\theta^k)\Big)$$

Has first- and second-order derivatives which are continuous over Θ, admits a global minimum $\hat{\theta}_n^k$ which belongs to Θ and almost surely converges and uniformly in θ^k to a function $W(\theta^k)$ which is in turn has first- and second-order derivatives and has a unique global minimum at $\theta_*^k \in \Theta$ such as $V_n''(\theta^k) \rightarrow W''(\theta^k)$ almost surely and uniformly in $\theta^k \in \Theta$ as $n \rightarrow +\infty$. Note that the preceding conditions imply that $\hat{\theta}_n^k$ converge almost surley to θ_*^k as $n \rightarrow +\infty$. We use in after the V_n function to build the modified Schwarz Criterion.

2.2 Schwarz Information Criterion for the incomplete data

2.2.1 Lemma

Consider two sequences of positive random variables (T_n) and (U_n) and a convergent positive sequence (α_n) defined as $[U_n > \alpha_n]$ implies$[T_n > U_n]$. Suppose there are two postive constants γ and ϵ such as:

$$P[(T_n - \alpha_n) \geq \gamma] \geq \epsilon \quad \forall n$$

Then

$$\exists N, \forall n > N, \forall \delta > 0 \qquad \ln E\Big(T_n^{\ln(n)}\Big) - \ln E\Big(U_n^{\ln(n)}\Big) > -\delta.$$

2.2.2 Proposition 2

Let (M_k) a sequence of models such that M_k describes the incomplete data. Let

$f\Big(.\Big|(M_k,\theta^k)\Big)$ And $h(Y)$ respectively the density of probability and the marginal density of Y . For each model candidateM_k, we associate the posterior probability $P(M_k|Y)$and

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301478

the prior probability$P(M_k)$. We consider $\hat{\theta}_n^k$ the estimator of θ^k obtained by maximizing the likelihood. We suppose that $\hat{\theta}_n^k \rightarrow \theta_*^k$ almost surely.

We define E_Y the expected value with respect to the density $f\left(Y \middle| (M_k, \theta_*^k)\right)$ and let:

$I_{oc}(\theta|Y_{obs})$=$E_{Y_{mis}}\left(-\frac{\partial^2 \ln f(Y|\theta)}{\partial\theta\partial\theta'}\right)$

$$I_o(\theta|Y_{obs}) = -\left(-\frac{\partial^2 \ln f(Y_{obs}|\theta)}{\partial\theta\partial\theta'}\right)$$

Then, there exist $n_o \in IN$ such that for $n > n_o$, we have the following inequality :

$$-\frac{2}{n}\ln P(M_k|Y) \leq \frac{2}{n}\ln h(Y) - \frac{2}{n}\ln P(M_k) - 2Q\left(\hat{\theta}_n^k \middle| \hat{\theta}_n^k\right) + \ln(n) trace\left\{I_{oc}\left(\hat{\theta}_n^k \middle| Y_{obs}\right) I_o^{-1}\left(\hat{\theta}_n^k \middle| Y_{obs}\right)\right\}$$

2.2.3 The Modified Schwarz Criterion

The dimension k_o of the fitted model obtained by maximizing the posterior probability$P(M_k|Y)$. We base on the derivation of the new criterion on the majorant of $-\frac{2}{n}\ln P(M_k|Y)$ represented in the proposition 2.

If we consider only terms which depend on k, the estimator $\hat{k}$ of the unkown order k_o is obtained with the minimum of the following quantity:

$$-\frac{2}{n}\ln P(M_k) - 2Q\left(\hat{\theta}_n^k \middle| \hat{\theta}_n^k\right) + \ln(n) trace\left\{I_{oc}\left(\hat{\theta}_n^k \middle| Y_{obs}\right) I_o^{-1}\left(\hat{\theta}_n^k \middle| Y_{obs}\right)\right\} \quad (1.1)$$

In this expression, we can eliminate the prior probability P (Mk) because when the integer k is lower or equal to the maximum order L, we consider $P(M_k) = \frac{1}{L}$

If k ∈ IN *, the choice of P (M_k) corresponds to the coding of integers. For example, since the optimal coding defined by Rissanen[12,13], we can write:$P(M_k) = \frac{1}{c}2^{-log^* k}$

Where log*k = log2k + log2log2k + ... and c is the constant of normalization such as:

$\frac{1}{C}\sum_{k=1}^{\infty} 2^{-log^* k} = 1$ (c ≈ 2.865064)

In the continuation, we suppose that 1 ≤ k ≤ L thus$P(M_k) = \frac{1}{L}$. Using (1.1) we propose the following criterion which we note SIC_{cd}:

$$SIC_{cd}(k) = -2Q\left(\hat{\theta}_n^k \middle| \hat{\theta}_n^k\right) + \ln(n)\, trace\{I_{oc}\left(\hat{\theta}_n^k \middle| Y_{obs}\right) I_o^{-1}\left(\hat{\theta}_n^k \middle| Y_{obs}\right)\}$$

$$\hat{k} = \underset{1 \leq k \leq L}{\operatorname{argmin}} SIC_{cd}(k)$$

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301478

3. SIMULATION STUDIES

In order to validate the theoretical results, we consider an autoregressive processes Y_t of order 3. We suppose that Y_t is causal and we take

$$Y_t + 0.653 * Y_{t-1} - 0.064 * Y_{t-2} - 0.227 * Y_{t-3} = \varepsilon_t$$

Where ε_t is the white noise with mean 0 and variance σ^2.

By the Software analyzer Splus we generate 500 samples of size n when we vary the values of size and variance in {50, 200} and {1, 2} respectively. We construct then the incomplete data by eliminating the observations of each sample, this operation is according to a discard probability P_{mis}; we denote that P_{mis} is the probability to remove some observations is set at 0.2, 0.33 and 0.4.

For each of the 500 incomplete data in a set, all parameter models in the candidate class are fit to the data using the EM algorithm. We calculate the order of models by the selection criteria SIC, AIC, AIC_{cd} and SIC_{cd}. We consider in this calculation the orders 1, 2, 3, 4 and 5.

We presented in four tables 1, 2, 3 and 4 the numerical results of criteria depending on the size of samples and the values of the variance.

We choose the variance $\boldsymbol{\sigma^2}$ = 1 for the tables 1 and 2 and we present the results for the variance $\boldsymbol{\sigma^2}$ = 2 in tables 3 and 4.

Table 1: Frequencies (%) of estimated orders, n = 200 and σ^2 = 1

Pmis	ORDER	CRITERIA			
		AIC	SIC	AICcd	SICcd
	1	00	2	00	2
	2	2.67	6	2.67	6
0	3	78.66	91.33	78.66	91.33
	4	14	0.67	14	0.67
	5	4.67	00	4.67	00
	1	00	2.66	00	2.70
	2	4	12.67	6.67	13.30
0.2	3	58.77	66	58.77	78.67
	4	24	12	24	5.33
	5	13.23	6.67	10.56	00
	1	00	2	00	2
	2	4	13.33	3.33	10.17
0.33	3	42.67	56.67	42.67	73.83
	4	24.67	18.67	28	4.97
	5	28.66	9.33	26	9.03
	1	5.33	15.22	1.46	10
	2	4	17.33	4	14.67
0.4	3	26	32.67	23.34	49.36
	4	30	23.34	36	19.30
	5	34.67	11.44	35.20	6.67

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301478

Table 2: Frequencies (%) of selected orders, n = 50 and σ^2 = 1

Pmis	ORDER	CRITERIA			
		AIC	SIC	AICcd	SICcd
	1	21	28	21	28
	2	10	14.67	10	14.67
0	3	39	54	39	54
	4	18	1.33	18	1.33
	5	12	2	12	2
	1	24	24	33	21.33
	2	10.67	14.67	12.33	12.67
0.2	3	29.33	36.67	32.67	44.67
	4	16	8.66	10	18.67
	5	20	16	12	2.66
	1	8.99	14	5	18
	2	9.01	26	12	26
0.33	3	26.64	30	27.66	36
	4	32	25.33	26.67	17.30
	5	23.36	4.67	28.67	2.70
	1	2	20	3.03	21.10
	2	12.87	13.33	11.22	17
0.4	3	16.60	20.67	14.78	26
	4	11.13	12	12	13
	5	57.40	34	58.97	22.90

Table 3: Frequencies (%) of selected dimensions, n = 200 and σ^2 = 2

Pmis	ORDER	CRITERIA			
		AIC	SIC	AICcd	SICcd
	1	2	3.34	2	3.34
	2	1.23	10	1.23	10
0	3	66.77	83.33	66.77	83.33
	4	22.54	3.33	22.54	3.33
	5	7.46	00	7.46	00
	1	00	7.30	3.30	8
	2	15.33	24	11.36	14
0.2	3	48	56	54.37	76
	4	26	6.70	24.97	2
	5	10.67	6	6	00
	1	00	6.67	00	9.66
	2	7.53	12.67	10.67	13.33
0.33	3	35.13	49.35	45.33	66.67
	4	26	19.31	20	3.67
	5	31.34	12	24	6.67
	1	7.33	18	4.64	12
	2	8	12	9.33	16.67
0.4	3	21.33	32	18	44
	4	30	28	39.36	20
	5	33.34	10	28.67	7.33

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301478

Table 4: Frequencies (%) of selected orders, n = 50 and σ^2 = 2

Pmis	ORDER	CRITERIA			
		AIC	SIC	AICcd	SICcd
	1	18.67	22	18.67	22
	2	14.67	15.34	14.67	15.34
0	3	49.33	58	49.33	58
	4	10	00	10	00
	5	7.33	4.66	7.33	4.66
	1	7.33	30	10	25.53
	2	20	17.31	16	18
0.2	3	24	42.70	27.33	44.47
	4	20.67	6.69	24	8
	5	28	3.30	22.67	4
	1	4.87	14	2	16.88
	2	13.53	14.67	10	11.12
0.33	3	19.13	34	22.67	38
	4	18.47	20	21.33	12
	5	44	17.33	44	22
	1	6.67	12.97	8	10
	2	0.57	5.03	4	20.67
0.4	3	19.33	28	16	32.60
	4	18.67	20	12.77	20.73
	5	54.76	34	59.23	16

In analyzing the results described in each of the four tables, we observe, generally, that the frequency of selection of the exact order k_o = 3(”good selection”) is a decreasing function in relation to the probability P_{mis} and the variance σ^2 of the white noise. We also observe that; the frequency of ”good selection” by the criterion SICcd is always superior to the one of the criteria AIC, SIC and AIC_{cd}, this result is independent to the size of the sample and the variance of the white noise.

The criterion SIC_{cd} is more performant than the other criteria; the frequency of ”good selection” of the order obtained through the criteria AIC and AICcd for 20% is similar to the one of SIC_{cd} with 40% of missing data when σ^2 = 2. Let's note otherwise that the decrease of the ”good selection” frequency by the criteria AIC, SIC and AIC_{cd} is faster than the one of criteria SIC_{cd}.

4. CONCLUSION

We have shown that the Schwarz criterion SIC, which is convergent for complete data, does not allow a correct estimation of the order when we have a meaningful number of missing data in the sample. We also verified by simulation that; the new criterion SIC_{cd} is stronger than criteria AIC, SIC and AIC_{cd} since the frequency of selection of the exact order by this new criterion is systematically superior to the one of selection obtained through the criterion AIC_{cd} proposed by Cavanaugh and Shumway [5] and generally with the classic criteria. The modified criterion of Schwarz SIC_{cd} presents therefore a particular interest in relation to the criteria information when we have the incomplete data.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 12-2023
DOI: 10.5281/zenodo.10301478

References

1) Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in second international symposium of Information Theory, ed. B. N. Petrov and F. Csaki, Akademia Kiado, Budapest, 267-281.

2) George E. P. Box and Gwillym M. Jenkins (1976), Time series analysis forecasting and control, revised edition, Holden-Day, page 274-284.

3) Peter J. Brocwell and Richard A. Davis (1991), Time series: Theory and methods, second edition, Springer-Verlag, page 282-286.

4) J. E. Cavanaugh, Andrew A. Neath (1999), generalizing the derivation of the Schwarz information, for communications in statistics-theory and methods. Volume 28, page 49-66.

5) J. E. Cavanaugh, Robert H. Shumway (1998), An Akaike Information Criterion for model selection in the presence of incomplete data, from Journal statistical Planning and Inference, Volume 67, page 45-65.

6) P. Dempster, N. M. Laird and D. Rubin (1977), Maximum Likelihood from Incomplete Data via the EM algorithm, Journal of the Royal Statistical Society.

7) G. J. McLachlan, Thriyambakam Krishnana, The EM algorithm and Extensions, A Wiey-Interscience Publication, JOHN WILLEY ET SONS, INC, 1997.

8) Schwarz, G. (1978), estimating the dimension of a model, The Annals of Statistics, Vol.6, 461-464.

9) Shibata, R. (1976), Selection of the order of an autoregressive model by Akaike's information criterion, Biometrika, 63, 117-126.

10) Shimodaira, H. (1994), a new criterion for selecting models from partially ob- served data, Springer-Verlag, New York, 21-29.

11) Rissanen J. (1981), Universal modeling and coding, IEEE, Trans. inf. theory, vol. IT27, 1, 12-23.

12) Rissanen J. (1983), a universal prior for integers and estimation by minimum description lenght, the annals of statistics, vol. 11, 416-431.

13) Xiao-Li Meng and Donald B. Rubin (1991), Using EM to obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm, volume 86, n416, Applications and case studies.