END
DATE
FILMED
9 82
DTIC

AD A118598

THE SELECTION OF SMOOTHNESS PRIORS FOR DISTRIBUTED LAG ESTIMATION

Hirotugu Akaike

DTIC
SELECTED
AUG 2 6 1982
F

82 08 25 152

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

THE SELECTION OF SMOOTHNESS PRIORS FOR DISTRIBUTED LAG ESTIMATION

Hirotugu Akaike*

Technical Summary Report #2394
June 1982

ABSTRACT

In the application of Shiller's smoothness prior for distributed lag
estimation the main difficulty is the selection of hyperparameters of the
prior distribution. In this paper the use of a maximum likelihood procedure
is proposed for this purpose and its performance is demonstrated by numerical
examples.

*The Institute of Statistical Mathematics, Tokyo, Japan

SIGNIFICANCE AND EXPLANATION

The distributed lag estimator developed by R. J. Shiller based on the concept of smoothness prior is a significant example that demonstrates the potential of Bayesian approach in statistics. However, the practical application of Shiller's estimator has been hampered by the difficulty of specifying the prior distribution.

In this paper an objective procedure for the selection of the prior distribution is proposed and its performance is checked by both artificial and real numerical examples. The result clearly demonstrates the practical utility of the procedure. It also shows the danger inherent in an arbitrary subjective choice of the prior distribution.

It is expected that the result reported in this paper will contribute to the development of practical applications of Bayesian statistics.

THE SELECTION OF SMOOTHNESS PRIORS FOR DISTRIBUTED LAG ESTIMATION

Hirotugu Akaike*

1. INTRODUCTION

Shiller (1973) introduced the concept of smoothness priors to define a Bayesian estimator of the lag coefficients, or the impulse response function, of a linear system. In this approach the prior preference of the smoothness of the lag function is expressed by a spherical normal distribution of the fixed order differences of the lag coefficients.

One outstanding problem in the application of Shiller's estimator is the choice of the lag length, the order of differencing and the variance of the spherical normal prior distribution. In this paper we propose a practical solution to this problem obtained by maximizing a properly defined likelihood of the Bayesian model.

The performance of the estimator is checked by numerical examples. The result shows that the estimator performs satisfactorily under widely varying condition. It also shows that an arbitrary subjective choice of the prior distribution can produce awkward estimate.

2. SMOOTHNESS PRIOR FOR DISTRIBUTED LAG ESTIMATION

Consider the stochastic linear system defined by

$$(1) \qquad y_n = \sum_{m=0}^{M} a_m x_{n-m} + w_n$$

where y_n, x_n and w_n denote the output, input and error term of the system, respectively. Here $\{w_n\}$ is assumed to be a sequence of random variables which are independent of $\{x_n\}$ and are independently identically distributed as normal with mean zero and variance σ^2.

*The Institute of Statistical Mathematics, Tokyo, Japan

Shiller introduced the smoothness prior defined by a spherical normal distribution of the dth differences of the distributed lag coefficients (a_m). This prior distribution is given by

$$p(a|\sigma,\lambda) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{M+1}{2}} |\lambda^2 R_d R_d'|^{1/2} \exp\left(-\frac{\lambda^2}{2\sigma^2} a' R_d' R_d a\right) ,$$

where $| \ |$ denotes the determinant, $a = (a_0, a_1, \ldots, a_M)'$, λ is the ratio of the standard deviation σ of w_n to that of the dth difference of a_m and $R_d = R_1^d$, where R_1 is an $(M + 1) \times (M + 1)$ matrix defined by

$$R_1 = \begin{bmatrix} 1 & -1 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & -1 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & 1 \end{bmatrix} .$$

It should be noticed here that the last d rows of R_d define incomplete differences. However, as was mentioned by Shiller, this is equivalent to connecting the final coefficients to $a_m = 0$ for $m > M$, which will be a reasonable assumption for practical application if the lag length M is taken sufficiently large.

By combining the present prior distribution with the data distribution

$$p(y|a,\sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2} |y - Xa|^2\right) ,$$

where $y = (y_1, y_2, \ldots, y_N)'$, $| \ |$ denotes the Euclidean norm, and

$$X = \begin{bmatrix} x_1 & x_0 & \cdot & \cdot & \cdot & x_{1-M} \\ x_2 & x_1 & \cdot & \cdot & \cdot & x_{2-M} \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ x_N & x_{N-1} & \cdot & \cdot & \cdot & x_{N-M} \end{bmatrix} ,$$

the estimate a^* of the lag coefficients is defined as the posterior mean of a and is

-2-

given by

$$a_e = (X'X + \lambda^2 R_d'R_d)^{-1}X'y .$$

Shiller demonstrated by numerical examples the superiority of this type of estimator to both the ordinary least squares estimator and the Almon lag estimator. This is one of the earliest examples of successful Bayesian modeling of practical importance.

The crucial point in applying Shiller's estimator is the choice of the hyperparameter λ. Shiller suggested some rule of thumb for the choice of λ. However, this is deeply concerned with the basic problem of the selection of a Bayesian model which is vital in implementing a Bayesian procedure.

3. SELECTION OF A PRIOR DISTRIBUTION BY LIKELIHOOD

When there are finite number of Bayesian models defined by the data distributions $f_k(\cdot|\theta_k)$ and corresponding prior distributions $\pi_k(\theta_k)$ ($k = 1,2,\ldots,K$) the posterior probability of each model is given by

$$p(k|y) = \frac{f(y|k)C_k}{\sum\limits_{k=1}^{K} f(y|k)C_k} ,$$

where y denotes the observation, C_k denotes the prior probability of the kth model and $f(y|k)$ is defined by

$$f(y|k) = \int f_k(y|\theta_k)\pi_k(\theta_k)d\theta_k .$$

The above formula of $p(k|y)$ shows that it is natural to call $f(y|k)$ the likelihood of the Bayesian model specified by $f_k(\cdot|\theta_k)$ and $\pi_k(\theta_k)$.

Recent work by the present author suggests that when there is no further prior information available and the distributions $f(\cdot|k)$ are well separated, i.e., only one model attains high likelihood for one particular observation, then the equal prior probability $C_k = 1/K$ is a natural choice that let the data speak most (Akaike, 1982). With this choice of the prior probability distribution the maximum likelihood selection that chooses the Bayesian model with maximum $f(y|k)$ is equivalent to the selection by maximum posterior probability. This shows that under certain circumstances the selection

-3-

of a Bayesian model by maximizing the likelihood can be a reasonable procedure from the Bayesian point of view. In this paper we pursue the possibility of applying this idea to the problem of selection of the smoothness prior.

In practical applications of the smoothness prior for the distributed lag estimation we usually do not know the value of σ. Accordingly we have to specify a prior distribution of σ. To make the resulting estimator applicable to observations in arbitrary scale unit we consider the use of Jeffrey's ignorance prior $\sigma^{-1}d\sigma$. For each particular application we may consider a proper prior distribution $C(u,v)\sigma^{-1}$ obtained by restricting the range of σ to a finite interval (u,v) of positive numbers. However, since the integral of $p(y|a,\sigma)p(a|\sigma,\lambda)\sigma^{-1}$ with respect to $dad\sigma$ is finite, we may use this integral as the likelihood of the Bayesian model specified by $p(y|a,\sigma)$ and $p(a|\sigma,\lambda)C(u,v)\sigma^{-1}$ with sufficiently small u and sufficiently large v. Since only ratios of the likelihoods are of interest we define the likelihood of the Bayesian model, specified by the data distribution $p(y|a,\sigma)$ and the (improper) prior distribution $p(a,\sigma|\lambda) = p(a|\sigma,\lambda)\sigma^{-1}$, by

$$p(y|\lambda) = \iint p(y|a,\sigma)p(a,\sigma|\lambda)dad\sigma .$$

Since the hyperparameter λ has a clearly defined technical meaning as the ratio of the standard deviation of an error term to that of a difference of lag coefficients, it is not difficult to find a reasonable selection of finite number of possible values of λ. The model selection is realized by maximizing the likelihood over this set of λ.

For our present model we have

$$p(y|a,\sigma)p(a,\sigma|\lambda) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N+M+1}{2}} |\lambda^2 R_d'R_d|^{1/2}\frac{1}{\sigma}$$

$$\cdot \exp\left[-\frac{1}{2\sigma^2}(a - a_e)'(X'X + \lambda^2 R_d'R_d)(a - a_e)\right]\exp\left[-\frac{1}{2\sigma^2}s(\lambda)\right] ,$$

where $a_e = (X'X + \lambda^2 R_d'R_d)^{-1}X'y$ and $s(\lambda) = |y|^2 - a_e'(X'X + \lambda^2 R_d'R_d)a_e$. The likelihood $p(y|\lambda)$ of this model is then given by

$$p(y|\lambda) = \frac{1}{2} \left(\frac{1}{\pi}\right)^{\frac{M}{2}} \Gamma\left(\frac{M}{2}\right) s(\lambda)^{-\frac{M}{2}} |x'x + \lambda^2 R_d' R_d|^{-\frac{1}{2}} |\lambda^2 R_d' R_d|^{\frac{1}{2}} .$$

Obviously the estimator a_* is given as the solution of the least squares problem that minimizes $|y^* - x^*a|^2$, where

$$x^* = \begin{bmatrix} x \\ \lambda R_d \end{bmatrix} , \quad y^* = \begin{bmatrix} y \\ 0 \end{bmatrix} .$$

Here it holds that $s(\lambda) = |y^* - x^*a_*|^2$. By successively applying the Householder transformation to $[x^* y^*]$, first to transform λR_d into upper triangular form and then to transform the whole matrix into upper triangular form, the necessary quantities for the likelihood computation can easily be obtained during the process of the least squares computation. For the purpose of the comparison of models we may ignore the constant factor $2^{-1} \pi^{-\frac{M}{2}} \Gamma(M/2)$ of the likelihood.

It should be noted here that the likelihood $p(y|\lambda)$ is also a function of d, the order of differencing, and M, the lag length. Thus our search for the best model is realized by maximizing $p(y|\lambda)$ over some finite number of possible combinations of λ, d and M. The practical utility of this procedure will be demonstrated by numerical examples in the next section. The feasibility of this type of procedure was first discussed in Akaike (1980a) and its application to seasonal adjustment was discussed by Akaike (1980b). The idea of maximizing a likelihood with respect to the hyperparameter is discussed in an earlier paper by Good (1965).

4. NUMERICAL INVESTIGATION

From the point of view of data analysis a Bayesian model is simply an artificial construction that allows to generate an output from a given set of data. Only when we confirm that it often produces results superior to those obtained by other conventional procedures we can claim the usefulness of the Bayesian procedure. In this section we will

discuss the practical utility of the distributed lag estimation realized by the maximum likelihood selection of the smoothness prior.

First we discuss the application to the artificial example discussed by Shiller as his second example. In this example the input series was the series of four to six month commerical paper rate which gave a typical nearly collinear matrix X. The output series y was generated by the relation $y = Ax + w$ with $w = 0.05e$, where e is a vector of standard normal random numbers. The lag coefficients were defined by $a_m = \phi((\frac{1}{2})(m - 9))$ with $M = 19$, where $\phi(t)$ denotes the density of the standard normal distribution. The dimension, or the length, of y was given by $N = 40$. One data set generated by this model was used in the subsequent analysis.

The search for the smoothness prior was extended over the models defined with $\lambda = 5 \times 2^k$ $(k = -10,(1),10)$, $M = -1,(1),19$, where $M = -1$ denotes zero regression, and $d = 1,2,3$. Since we are accustomed to the use of log likelihood ratio test we used $(-2)\log p(y|\lambda)$ as our criterion. By ignoring the additive constant the criterion is given by

$$ABIC = N \log s(\lambda) + \log|X'X + \lambda^2 R_d' R_d| - \log|\lambda^2 R_d' R_d| \, ,$$

where N is the dimension of the vector y, log denotes natural logarithm and ABIC stands for a Bayesian information criterion (Akaike, 1980a). Our search for the model was realized by finding the minimum of ABIC.

The best choice in terms of the criterion was given by $M = 13$, $d = 2$ and $\lambda = 5 \times 2^{-3}$. The resulting estimate is shown in Table 1 along with the true values and the ordinary least squares estimate for $M = 19$, denoted by LS. By Table 1 it is obvious that the present estimation procedure is producing significantly improved estimate over the ordinary least squares estimate. Although the present result is not directly comparable with Shiller's result due to the use of different realizations of the error term, the shape of the estimated distributed lag coefficients is quite similar to some of the best ones given by Shiller.

One might be concerned with the possibility of non-smooth behavior of distributed lag coefficients. To check the performance of our procedure under such circumstances we

-6-

TABLE 1

RESULTS OF AN EXPERIMENT OF SHILLER'S SECOND EXAMPLE.
LS DENOTES LEAST SQUARES ESTIMATE.

m	0	1	2	3	4
True	.000	.000	.001	.004	.018
a*	-.011	-.001	.008	.014	.015
LS	-.010	.021	-.045	.037	.078

m	5	6	7	8	9
True	.054	.130	.242	.352	.399
a*	.045	.136	.240	.347	.400
LS	-.074	.255	.113	.462	.334

m	10	11	12	13	14
True	.352	.242	.130	.054	.018
a*	.370	.250	.125	.060	0
LS	.359	.329	.046	.072	.042

m	15	16	17	18	19
True	.004	.001	.000	.000	0
a*	0	0	0	0	0
LS	-.018	-.050	.065	-.008	-.008

considered a system defined by

$$y_n = 1.2\, x_n - 0.6\, x_{n-1} + 0.4\, x_{n-2} + w_n \, ,$$

where the input x_n was the same as that in the preceding example and thus $N = 40$, and w_n was also normal with mean zero and $\sigma = 0.05$. The range of the search for the parameters was the same as in the preceding example. The minimum of ABIC was attained at $d = 1$, $M = 2$ and $\lambda = 5 \times 2^{-7}$. The resulting estimate a_* and the estimate a_{**}, which was obtained by the parameters used for the computation of a_* of the preceding example, are given in Table 2 along with the true values and LS, the ordinary least squares estimate for $M = 19$.

The most remarkable finding with this result is that our procedure produced extremely good result. This was made possible by the correct determination of M and the choice of a very small value as λ. Contrary to this, a_{**} which was obtained by the parameters of Table 1 produced very poor result, even worse than LS. This clearly demonstrates the danger of applying a Bayesian model based on an arbitrary choice of the prior distribution. It is obvious that a proper procedure of adaptation is necessary for the practical application of the smoothness prior.

Having confirmed the performance of our procedure at the two extreme situations, most favorable and unfavorable, we now turn to the example of real data handled by Shiller as his first example. In this example Shiller analyzed the response of the Federal Reserve Board Aaa new issue yield series to the four to six month prime commercial paper rate. Due to the unavailability of the Federal Reserve Board Aaa new issue yield series we used the corresponding series of Moody's AAA bond yield as the output series. The similarity of our result to Shiller's confirms that the substitution did not change the essential aspect of the problem.

In this example it is already noticed by Shiller that the coefficient a_0 has different characteristic from other coefficients and should be freed from the rest. The validity of this hypothesis can easily be checked by our present approach by considering models obtained by multiplying the first row of R_d by a small positive number δ. For the present example ABIC kept decreasing when δ was decreased from 1 to 0.0001 in

-8-

TABLE 2

TEST RESULT FOR NON-SMOOTH DISTRIBUTED LAG COEFFICIENTS.
a** DENOTES THE ESTIMATE OBTAINED BY THE PARAMETERS OF a* OF TABLE 1.

m	0	1	2	3	4
True	1.2	-.6	.4	0	0
a*	1.206	-.601	.394	0	0
a**	.859	.068	-.075	.051	.090
LS	1.193	-.587	.390	-.044	.045

m	5	6	7	8	9
True	0	0	0	0	0
a*	0	0	0	0	0
a**	.037	-.025	-.012	.021	-.010
LS	.018	-.018	.037	-.037	.026

m	10	11	12	13	14
True	0	0	0	0	0
a*	0	0	0	0	0
a**	-.000	.039	-.000	-.038	0
LS	-.029	.030	-.027	.059	-.067

m	15	16	17	18	19
True	0	0	0	0	0
a*	0	0	0	0	0
a**	0	0	0	0	0
LS	.028	-.015	-.052	.092	-.043

several steps. This confirmed the validity of Shiller's argument. The following result was obtained with $\delta = 0.0001$ and the same range of the parameters as in the preceding examples.

The minimum of ABIC was attained at $d = 3$, $M = 19$ and $\lambda = 5 \times 2^5$. The amount of reduction of the minimum ABIC obtained by reducing δ from 1 to 0.001 was 20.7. The estimate $a_\bullet$ is given in Table 3 along with the estimate $a_\bullet^{1.0}$ obtained by putting $\delta = 1.0$. In this example the constant term a_{-1} was included to represent the effect of non-zero average values of the input and output series. The estimate denoted by $a_\bullet^d$ was obtained from the series of first order differences of the input and output series. This was to check the possible effect of the trends of both series. The estimate $a_\bullet^d$ was obtained with $d = 3$, $M = 19$ and $\lambda = 5 \times 2^4$. The corresponding ordinary least squares estimate for the differenced series is given by LS.

The similarity between $a_\bullet$ and $a_\bullet^d$ is remarkable and confirms that the smooth behavior of $a_\bullet$ does not represent the spurious response due to the trend components. Also the distortion caused by the inclusion of a_0 can be seen clearly from $a_\bullet^{1.0}$. The usual erratic pattern of the least squares estimate persists in this example.

Summarizing our observations of numerical results, including those not reported here, we may conclude as follows:

1) the estimator is most sensitive to the choice of λ,

2) the choice of M, the lag length, is also fairly critical,

3) the choice of d, the order of differencing, is not so critical.

We also found that the selection of M must be done with ABIC minimized with respect to λ. Without the adjustment of λ the selection of M by minimizing ABIC produced poor results.

RESULT OF APPLICATION TO REAL ECONOMIC DATA. $a_0^{1.0}$ DENOTES THE RESULT OBTAINED UNDER THE ASSUMPTION OF SMOOTH CONNECTION TO a_0. a_*^d DENOTES THE RESULT OBTAINED FROM DIFFERENCED SERIES.

m	0	1	2	3	4
a^*	.212	.029	.045	.058	.068
$a_*^{1.0}$	.104	.088	.076	.070	.068
a_*^d	.281	.005	.031	.052	.068
LS	.352	-.132	.081	.127	-.009

m	5	6	7	8	9
a^*	.074	.079	.078	.076	.072
$a_*^{1.0}$	.069	.072	.075	.076	.075
a_*^d	.078	.083	.084	.081	.076
LS	.106	.106	.084	.054	.144

m	10	11	12	13	14
a^*	.066	.058	.049	.040	.031
$a_*^{1.0}$	.072	.065	.056	.046	.035
a_*^d	.068	.059	.049	.039	.030
LS	.077	.058	.057	.025	.038

m	15	16	17	18	19
a^*	.022	.014	.008	.004	.001
$a_*^{1.0}$	.024	.014	.007	.002	0
a_*^d	.022	.014	.008	.004	.001
LS	.079	.048	.063	-.042	.037

5. DISCUSSION

The purpose of the present paper has been to show the feasibility of an objectively defined procedure for the selection of the smoothness prior for the estimation of distributed lag coefficients. It was confirmed that the maximum likelihood procedure proposed in this paper can poduce results cmparable to the results reported by Shiller in his original paper. Since Shiller's results may be considered as typical examples produced by an expert this shows that the present procedure is producing a good approximation to the judgement procedure of an expert. The next step of the Bayesian modeling will be the specification of a prior distribution for the lag length M.

The result of Table 2 demonstrated the robustness of the present estimation procedure. However, it also demonstrated the danger inherent in the purely subjective selection of a prior distribution.

Although the result reported in this paper has demonstrated the feasibility of the smoothness prior selection for distributed lag estimation, the practical applicability of the single input single output model to the analysis of economic data is rather limited. This is due to the common existence of feedback between the input and output in econometric applications. In such a case multivariate time series modeling is required. Whether the smoothness prior can find a useful application in this case is a subject of further study.

REFERENCES

Akaike, H. (1980a). Likelihood and the Bayes procedure. In _Bayesian Statistics_.

    (J. M. Bernardo, M. H. De Groot, D. V. Lindley and A. F. M. Smith, eds.). Valencia,

    Spain: University Press, pp. 143-166.

Akaike, H. (1980b). Seasonal adjustment by a Bayesian modeling. _Journal of Time Series_

    Analysis, 1, 1-13.

Akaike, H. (1982). On minimum information prior distributions. Tech. Rep. No.   ,

    Mathematics Research Center, University of Wisconsin-Madison.

Good, I. J. (1965). _The Estimation of Probabilities_. Cambridge, Massachusetts:

    M.I.T. Press

Shiller, R. J. (1973). A distributed lag estimator derived from smoothness priors.

    Econometrica, 41, 775-788.

HA/ed

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER 2394	2. GOVT ACCESSION NO. AD·A118598	3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle) THE SELECTION OF SMOOTHNESS PRIORS FOR DISTRIBUTED LAG ESTIMATION	5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
	6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s) Hirotugu Akaike	8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041

9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of 610 Walnut Street Wisconsin Madison, Wisconsin 53706	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics and Probability

11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709	12. REPORT DATE June 1982
	13. NUMBER OF PAGES 13

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)

18. SUPPLEMENTARY NOTES

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

Smoothness prior; Bayesian model; Likelihood; Hyperparameter;
Distributed lag; Impulse response

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

In the application of Shiller's smoothness prior for distributed lag estimation the main difficulty is the selection of hyperparameters of the prior distribution. In this paper the use of a maximum likelihood procedure is proposed for this purpose and its performance is demonstrated by numerical examples.

-8

DTIC