NBER WORKING PAPER SERIES

Alternative Prior Representations of 'Smoothness' for Distributed Lag Estimation

Robert J. Shiller*

Working Paper 89

COMPUTER RESEARCH CENTER FOR ECONOMICS AND MANAGEMENT SCIENCE
National Bureau of Economic Research, Inc.
575 Technology Square
Cambridge, Massachusetts 02139

June 1975

*NBER Research Fellow, 1974-75, and University of Pennsylvania.

Abstract

In some applications of the distributed lag model, theory requires that all lag coefficients have a positive sign. A distributed lag estimator which provides estimated coefficients with positive sign is developed here which is analogous to an earlier distributed lag estimator derived from "smoothness priors" which did not assure that all estimated coefficients be positive. The earlier estimator with unconstrained signs was a posterior mode of the coefficients based on a spherically normal "smoothness prior" in the d+1 order differences of the coefficients. The newer estimator with constrained sign is a posterior mode of the logs of the coefficients based on spherically normal "smoothness prior" on the d+1 order differences of the logs of the coefficients. The meaning of both categories of prior is discussed in this paper and they are compared to prior parameterizations of the lag curve. Both varieties of "smoothness prior", in contrast to the parameterizations, allow the coefficients to assume any "smooth" shape subject to the sign constraint. The sign-constrained estimator has the additional advantage that it easily forms asymptotes. Moreover, the sign constrained estimator is easily implemented. The estimate can be obtained by an iterative procedure involving regressions with dummy observations similar to those used to find the unconstrained sign estimator. An illustrative example of the application of both estimators is given at the end of the paper.

Acknoledgements

The author is indebted to Stanley Wasserman who wrote the TROLL Macro which implements the estimator developed here and which produced the illustrative example in this paper. This paper also benefited from discussions with Richard Hill, Edward Leamer and Forrest Nelson.

Contents

I. Introduction

Linear distributed lags are widely used in econometrics to model relationships between economic variables when the relationships are not well described in terms of a simple correlation between contemporaneous values of the variables but are rather distributed over time. Typically the response of an economic variable y to variations in another economic variable x may be sluggish and delayed. For instance, a response in terms of personal consumption expenditure (y) to government policy which influences disposable personal income (x) may be felt for a period of years following the policy action. In the absence of a theoretical structure for the relationship, it is often valuable to assume that it has a simple linear form.

A time series y_t at time t is said to follow a "linear distributed lag" on another time series x_t if:

$$y_t = \sum_{i=o}^{\lambda-1} \beta_i x_{t-i} + \varepsilon_t \qquad (1)$$

where the β_i are constant coefficients (which will have to be estimated) λ is the "lag length" and ε_t is a stationary stochastic process with zero mean and independent of x. In this paper we will be concerned with the question: how can we represent our prior knowledge concerning the vector β of distributed lag coefficients β_i i=0 λ-1? Since the distributed

lag model has many parameters, the representation of this prior knowledge is an essential element of our modeling. A number of parametrizations or prior distributions has been suggested for economic applications and we will discuss their meaning. In particular, we will discuss a prior distribution on β which we call "normal smoothness priors" (Shiller [1973]) and will develop a variant of it which we call "log normal smoothness priors". These priors can be used to derive estimators of the coefficients which have desirable properties. The new prior yields an estimator which has the property that all the β_i, $i=0, \lambda-1$, are specified to be greater than zero. The earlier prior yields an estimator with unconstrained sign.

We will also present here an illustrative example of an estimation problem using the constrained sign estimator as well as the unconstrained sign estimator.

II. The Estimation Problem and Traditional (Parametric) Approaches to it

The literature on the estimation of distributed lags is very extensive since it is a fundamental problem for econometric modelling. We cannot do justice to the litrature here. The reader is referred to two articles which survey the literature: Griliches [1965] and Sims [1974]. However, we do wish to make some comparison between the traditional parameterizations and a Bayesian approach with regard to their assumed prior knowledge.

The term "distributed lag" in economics we first coined by Irving Fisher in the 1920's who used this model to represent the response of interest rates to inflation rates. The distributed lag model may also be called a "linear transfer function" model or "linear filter" model. Such linear model have been used extensively in many branches of science and engineering. What distinguishes our econometric problem from these other applications is just the nature of the vague prior knowledge concerning the coefficients coupled with the shortage of data available in most economic applications.

When the relation (1) is structural (i.e. will continue to hold even after policy makers interfere with x) then we generally expect the lag coefficients β_i to trace out (if plotted against i) a "simple" or "smooth" curve. Given that (1) is a structural relation, then we can assess our prior beliefs regarding the β_i by asking what would we expect about the result of an experiment in which x_o is given a unit shock by policy makers without changing x in other time periods. The increment in y_t caused by the shock will be β_t. If the structural relation from x to y is a sluggish one, we would expect the curve β_t will be a smooth one which

tails off for high t. We may not, however, wish to rule out the possibility that the curve may be bimodal or have negative values over some interval, etc. The minimal kind of prior knowledge that we generally will wish to assume is that the curve is probably fairly simple.

Sims (1974) has emphasized that if a random economic time series y_t is projected on current and lagged values of another time series x_t, there is no reason to expect a "smooth" distributed lag even if the time series themselves are smooth. This should caution those who estimate distributed lags in cases in which there is no theoretical structure. However, cases in which (1) is likely to be useful for policy purposes are likely to be confined to those cases in which (1) in a relationship with a simple, stable form.

The problem is to estimate β_i i=0, λ-1 given n observations of y, $Y=\{y_{t-n+1}, y_{t-n+2}, \ldots . y_t\}$ and given a matrix X whose columns are current and lagged values of x, $X_{ij}=x_{t-n+i-j+1}$. Then we may write:

$$Y = X\beta + \varepsilon \tag{2}$$

where β is the vector $\{\beta_o, \beta_1, \ldots \beta_{\lambda-1}\}$ and ε is spherically normally distributed with precision h (i.e. with variance $\sigma^2=1/h$) and is independent of x. The likelihood function of β is then:

$$\ell(\beta|X,Y) \propto h^{n/2} \exp\left(-\frac{1}{2} h(Y-X\beta)'(Y-X\beta)\right) \tag{3}$$

The maximum likelihood estimate of β is then the ordinary least squares estimate $\hat{\beta}_{OLS}=(X'X)^{-1}X'Y)$.

The problem with the maximum likelihood estimate is that it is not based on any (proper) prior knowledge regarding the coefficients. Since economic time series are usually short and X exhibits multicollinearity the problem is significant. If the values β_i are plotted against i, the shape one typically gets in jagged and erratic, with coefficient frequently changing sign.

The simplest way to represent the assumption that the coefficents lie on a smooth curve is to parametrize the β_i, $\beta_i=g(i,a,b,c,..)$ and to reduce the problem to the estimation of a smaller group of parameters a, b, c,. The most popular such parametrizatiοn is that proposed by Almon [1965], in which the β_i are assumed to lie on d^{th} degree polynomial in i. Thus, $\beta=A\gamma$ where A is a matrix with d+1 independent columns whose i^{th} element is given by a polynomial in i of degree d or lower. The constrained maximum likelihood estimate $\hat{\beta}_a$ of β is then $\hat{\beta}_a=A(A'X'XA)^{-1}A'X'y$. The estimates of β_i will if plotted against i then lie on the familiar class of polynomials (a straight line if d=1, a parabola if d=2, etc.). While the coefficients are assured a more simple shape, they are probably constrained too much by this procedure. For instance, the coefficients cannot then "tail off". A polynomial can approximate an asymptote in a region only by "oscillating" around it, the number of slope changes limited to d.

The Almon procedure can easily be improved by substituting a different matrix A. Any parametrization of β which is linear in the parameters i.e. $\beta=A\gamma$ can be estimated as easily. For instance, piecewise linear distributed lags have been thus implemented. An improvement on the Almon

procedure would clearly be to assume an A matrix which can more easily produce asymptotes. We might hypothesize an A matrix whose columns consist of negative powers of i, for instance.

While such linear parametrizations may be convenient and useful, they have a limited ability to represent prior information. By setting $\beta=A\gamma$ where A has m independent columns, we constrain the coefficients β to lie on an (m^{th}) order linear manifold. There is no linear manifold which well represents the kind of prior knowledge we have regarding the coefficients. This kind of constraint always allows the coefficents β_i to change sign m-1 times or to fit m-1 "outliers". We cannot prevent erratic shapes without keeping m very small, in which case the lag has been constrained to lie on an elementary class of shapes.

Other parametrizations which are nonlinear have been suggested which are less convenient computationally but are in some cases more accurate as representations of prior knowledge. These include Koyck's geometric lag and its generalization the rational lag of Jorgenson which has a rational generating function, and Solow's negative binomial distributed lag.

Some of these parametrizations (e.g. the rational lag) may create problems because they do not prevent certain improbable lag structures and make assumptions about the tail of the lag distribution (for which the data may carry no information) which do not have any intuitive economic interpretation. When one uses the rational lag parametrization it is often found that the estimated lag oscillates in sign or explodes rather than tails off for high i. Before using any parametrization, one must always ask what the constraints mean in economic terms and whether these constraints are appropriate for the application.

III. Bayesian Approaches to Distributed Lag Estimation

An alternative approach to distributed lag estimation is a Bayesian approach which uses probabilistic prior information rather than parametrization (see e.g., Leamer [1972], Cleveland [1974], Shiller [1973]): This approach has the advantage that we do not impose arbitary constraints which we do not believe in. A Bayesian estimation procedure generally allows the estimated coefficients to take on any form, the prior beliefs only influence the estimates rather than constrain them and allow the data to overwhelm the priors in any direction in which the data is strong.

We are interested here in Bayesian priors which represent only the belief that the β_i trace out a "smooth" curve, and which carry no information about the size of any one coefficient considered separately. One class of such priors which was suggested in an earlier paper (Shiller, [1973]) we call, "Smoothness priors". The priors are designed to effectively break the near multicollinearity among the lagged variables that often produces erratic coefficient estimates even when the standard error of the regression in small, yet otherwise carry no information. These priors put a zero mean spherically normal distribution on the d+1 order difference u of the coefficients:

$$u = R\beta$$

where R is a $(\lambda-d-1 x \lambda)$ matrix which forms d+1 order differences of the coefficients and which has rank $\lambda-d-1 \equiv P$. The ij^{th} element of R is zero if $j-i>d+1$ or if $i>j$ and is otherwise equal to $(-1)^{j-i}\binom{d+1}{j-i}$. Given this prior on u, we take an uninformative prior $f(\beta_i)$= constant on any d+1 of the β_i, and get with a change of variables a prior which is uninformative on any of the β_i taken separately:

$$f(\beta) \propto \exp \frac{-1}{2\xi^2} \beta'R'R\beta \qquad (4)$$

If the prior variance ξ is small, then the priors assert that the coefficients will in some sense "hang together". In the limit, as ξ goes to zero, our priors approach the Almon prior that the coefficients lie on a d degree polynomial[2]. It is important to emphasize, however, that the priors are not well described is asserting that the coefficients lie near a d degree polynomial. One could have alternatively assumed, as did Maddala (1974), that $\beta_i = a_o + a_1 i + \ldots a_d i^d + \varepsilon_i$ where ε_i is spherically normally distributed and the coefficients $a_o, a, \ldots . a_d$ are independent of ε: and have a flat marginal distribution. The Maddala prior would assert that the coefficients can not deviate far from some polynomial and is indifferent as to how irregular are the deviations of the coefficients from the polynomial. The smoothness prior, on the other hand asserts, if ξ is relatively small, that the β_i can deviate dramatically from any polynomial if it does so gradually, i.e., in a "smooth" manner. If d=o then the priors will readily allow the lag curve to assume any shape which does not require the adjacent coefficients to be much different. If d=1 the priors allow any shape in which the slopes do not change quickly, i.e. it does not like "jagged" shapes. These may be understood as "flexible curve" priors. A flexible curve is a rubber ruler used by draftsmen to interpolate points. In a sense, the first degree smoothness prior allows any distributed lag shape which could easily be drawn using a flexible curve, that is, which does not require that the curve be bent too hard.

This prior, when applied to the likelihood function (3) yields a posterior by Bayes Law[1], assuming for the moment h is given, which is multivariate normal:

$$f(\beta|X,Y,h) \propto \frac{h}{2}^{\frac{n}{2}} \exp\left(-\frac{1}{2}h[(Y-X\beta)'(Y-X\beta)+K^2\beta'R'R\beta]\right) \tag{5}$$

where $K=\frac{\sigma}{\xi}$. One may take the mean (or mode) $\hat{\beta}_u$ as an estimate of β. This will be our estimator with unconstrained sign. The estimate $\hat{\beta}_u$ can be obtained by regressing Y and X matrices augmented with dummy observation. Defining:

$$\tilde{X} = \begin{bmatrix} X \\ KR \end{bmatrix} \quad \tilde{Y} = \begin{bmatrix} Y \\ 0 \end{bmatrix} \tag{6}$$

then the posterior mean can be found by regressing the augmented matrix $\tilde{Y}$ on the augmented matrix $\tilde{X}$:

$$\hat{\beta}_u = (\tilde{X}'\tilde{X}+K^2R'R)^{-1}\tilde{X}'\tilde{Y} \tag{7}$$

The great advantage of this procedure over parametrizations is most evident in cases in which the standard error of the regression is small and the X matrix exhibits near-multicollinearity. In this case, ordinary least squares will either fail altogether to produce a unique estimate or will produce a jagged erratic estimate. A parametrization will of course always produce a lag curve which lies in the class of elementary shapes that the parametrization allows, even if this produces a much higher standard error. The smoothness priors estimate, on the other hand, effectively deals with the multicollinearity by smoothing the curve,

but at the same time allowing the curve to take on any simple shape. If the true lag curve is a simple curve which does not lie near the class of curves specified by the parametrization estimates, than the smoothness estimate will have a much better fit. Moreover, the estimate $\hat{\beta}_u$ could not have been seen at all in the ordinary least squares estimate. One cannot visually "smooth" the ordinary least square estimate to produce a rough smoothness prior estimate, since in so doing one would not be taking into account the nature of the multicollinearity in the X matrix. These properties of the estimators as compared with the Almon estimator are illustrated in a case with a known lag curve in Shiller, [1973] and in Wilson [1975].

The smoothness prior estimate with unconstrained sign $\hat{\beta}_u$ has proven very useful, but suffers from a couple of problems at least in certain applications: 1. it is often difficult to specify the parameter K is not unit free and 2. the prior allows coefficients to change sign, whereas in many applications we believe the coefficients should all be positve or all be negative.

The first problem, that of choosing K, has led some authors to a ridge regression approach to the problem: Hill and Johnson [1975], Leamer [1974] and Maddala [1974]. It is true that the kind of prior information we have in applications of distributed lags may indeed be of the same vague nature as that which many think justifies the kind of judgemental approach inherent in the ridge regression procedure. The difference between our estimator $\hat{\beta}_u$ and the original ridge estimator is then merely that our priors relate to the differences of

the coefficients rather than their levels.

It is quite possible, on the other hand, that we can in fact easily specify a proper prior on some function of the coefficients which is unit free, such as their ratio. Even though we may have no prior notion as to the magnitude of the difference $\beta_i - \beta_{i+1}$ we may have prior information that, say, β_i should not differ from β_{i+1} by more than n%.

The natural extension of smoothness priors to deal with these problems is then a prior on the d+1 order difference V of the logs of the coefficients, (as mentioned in Shiller [1973]):

$$V = Rb$$

where $b = \log(\beta)$.

We then give V a spherically normal distribution with zero mean. If we now choose flat priors on d+1 of the $b_i = \log \beta_i$ = constant, then we get, with a change of variables, a prior on the b which is uninformative on any b_i considered separately:

$$f(b) \propto \exp(-\frac{1}{2\xi^2} b'R'Rb) \tag{9}$$

This expression, with a change of variables, implies a prior distribution on β which is a partially degenerate (uninformative) multivariate log normal distribution. The marginal prior on any β_i considered separately is the Jeffreys (1961) uninformative prior $f(\beta_i) \propto 1/\beta_i$. If R is a matrix which forms first differences (i.e. d=0) then the priors assert that, in effect, any lag shape is probable for which the proportional change between adjacent coefficients is not too high.

If ξ is very small, the priors reduce to the zero degree Almon contraint. If d=1, then the priors assert that the rate of change between adjacent coefficient should not change too fast, thus the prior also asserts that the lag curve should not be too "jagged". If d=1 then as $\xi \to 0$ the priors approach the Koyck constraint that the coefficients should lie on a geometric distribution. If d=2, the limiting case as $\xi \to 0$ is the constraint that the lag curve be proportional to a normal density (or its inverse). These limiting constraints are likely to be more acceptable than the polynomial constraints.

The log smoothness priors have the additional property that the prior conditional variance of β_i given adjacent coeficients β_{i-1}, β_{i+1}, etc., is small when the adjacent coefficients are small and large when the adjacent coefficients are large. The priors thus "tighten" up in regions in which the coefficients are small, as in the tail of the lag distribution; and assert that a single large coefficient in the tail is very improbable. The priors thus embody essentially what Leamer [1972] has called the "principle of proportionality" in distributed lag priors.[3] In contrast the variance of β_i conditional on β_{i-1}, β_{i+1},... is independent of β_{i-1}, β_{i+1},... with priors which are multivariate normal on the coefficients themselves rather than their logs.

Combining (9) with (3) and substituting e^b for β, we get, by Bayes Law, the posterior of b:

$$f(b|X,Y) \propto h^{n/2} \exp(-\frac{h}{2}[(Y-Xe^b)'(Y-Xe^b)+ K^2 b'R'Rb]) \qquad (10)$$

which is, unfortunately, not an easy distribution to deal with.

A modal estimate of b may, however, be found with an iterative procedure. We can write an expression which give the posterion mode $\hat{b}_c$ implicitly. It will be convenient to write the expression in terms of $\hat{\beta}_c \equiv \exp(\hat{b}_c)$ which will be our constrained sign estimate. By differentiating (10) with respect to b, setting to zero and substituting we get an implicit function for the mode $\hat{b}_c$. Substituting $\hat{b}_c = \log \hat{\beta}_c$

$$X'X \hat{\beta}_c + K^2 \text{ diag } (\hat{\beta}_c)^{-1} R'R \log \hat{\beta}_c = X'Y \qquad (11)$$

where K is σ/ξ and the matrix diag $(\hat{\beta}_c)$ is defined as a diagonal matrix whose iith element is $\hat{\beta}_i$. If the ordinary least squares estimate is positive, then as $K \to 0$ the estimate $\hat{\beta}_c$ approaches the ordinary least squares estimate. This property of the estimator of β is the result of having chosen as an estimate the mode of the posterior distribution of b rather than of the posterior distribution of β. In general, the limit of $\hat{\beta}_c$ as K goes to zero is the constrained maximum likelihood estimator in which all elements are forced to be positive.

An understanding of the behavior of the estimator is facilitated by considering the isodensity contours of the prior distribution of β and b. The simplest case, in which β has only two elements and when d=0 is shown in Figure 1. In Figure 1a, the isodensity for normal smoothness priors appear as a series of parallel lines. The center line, representing the prior mode, is a 45° line which passes through the origin. In Figure 1b we see the isodensity curves of the log normal smoothness priors, i.e., the prior on $b=\log(\beta)$, but expressed

in terms of β rather than b. These are a series of straight lines in the positive quadrant only which converge at the origin. Leamer [1975] has discussed both classes of isodensity contours; the first he calls "cylindrically uniform priors" and the second "conically uniform priors". A discussion of these contours is of course more general than a discussion of the prior distribution since more than one prior density can have the same set of contours.

In each case, the mode of the posterior distribution will lie on the locus of tangencies of the isodensity contours with isolikelihood contours. This locus has been called the "information contract curve" [Leamer, (1975)] or "curve decolletage". Just where along the curve the mode occurs depends on K.

The likelihood contours are concentric ellipses centered on the maximum likelihood estimate. In Figure 1, these are drawn for a (somewhat pathological) case in which the maximum likelihood estimate would make β_1 negative and β_2 positive. For Figure 1a the curve of tangencies is a straight line connecting the maximum likelihood estimate with the tangency of an isolikelihood ellipse with the center line of the isodensity contours. Modes can occur only on the segment (shaded darker) connecting the maximum likelihood estimate to the 45° line. The higher the value of K, the closer the Bayesian estimate will be to the 45° line. for low K, the estimate of β_1 is still negative, but for K sufficiently tight both coefficients must lie in the positive quadrant.

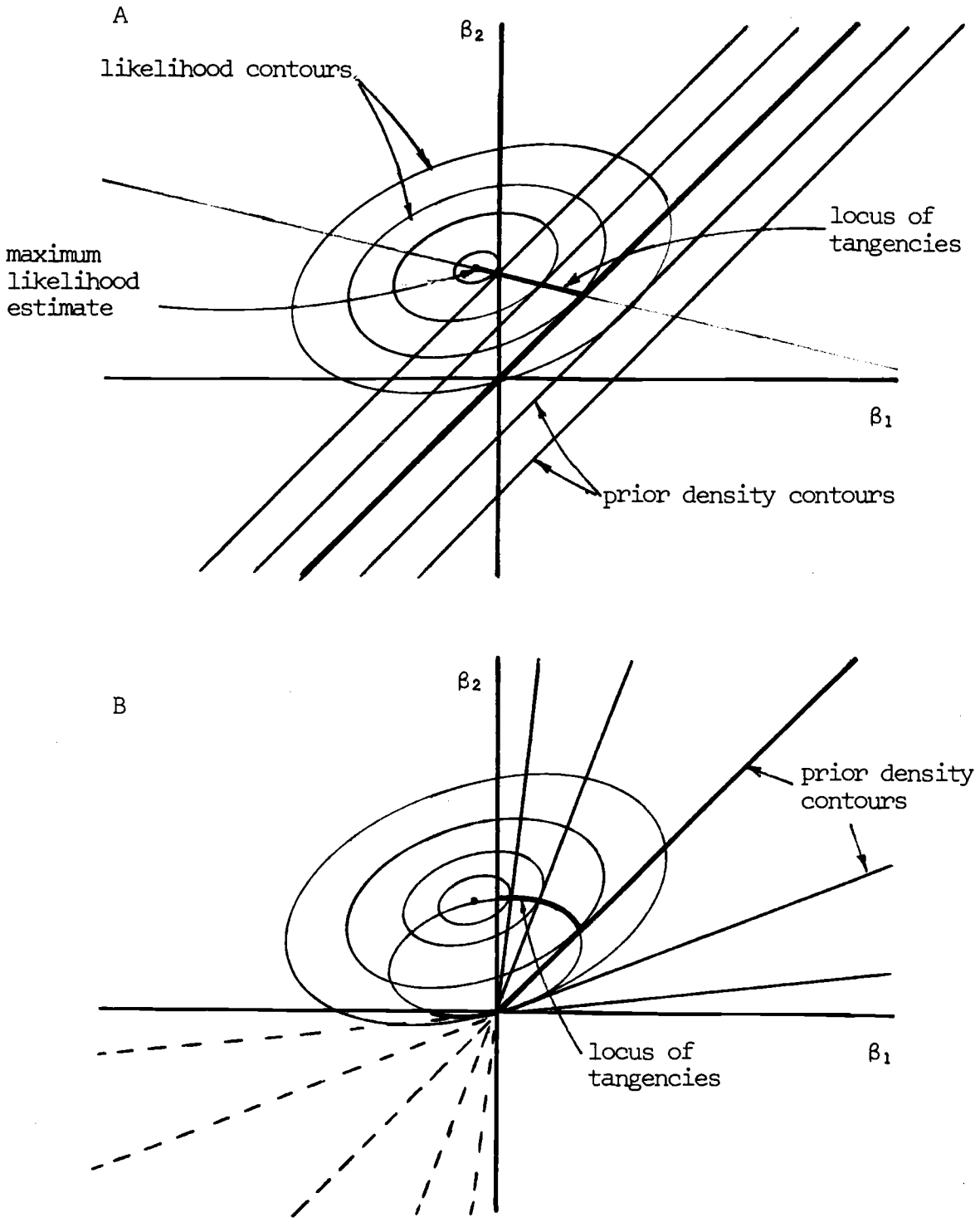

Figure 1. Density contours for smoothness priors, likelihood contours and their tangencies, A: normal smoothness priors, B: log normal smoothness priors.

For log normal smoothness priors, on the other hand, the locus of tangencies lies on an ellipse which passes through the maximum likelihood estimate and the origin, and which has the same eccentricity as the likelihood ellipses (for proof, see Leamer [1975]). Not all points on this ellipse are eligible modes however; only that part (which is shaded dark) which likes in the positive quadrant between the maximum likelihood estimate and the 45° line qualifies. As K approaches zero the estimate approaches the constrained maximum likelihood estimate (i.e., the tangency of a likelihood ellipse with the vertical axis).

As K approaches infinity, the estimate approaches the 45° line. For this case, when d=0, both estimators have the same limit as $K \to \infty$, but this result does not hold with higher d.

Had we specified the same prior on the ratios of the coefficients but instead required the individual coefficients to be negative rather than positive, we would find a mode which lies instead on the section of the tangency ellipse which lies in the negative quadrant between the 45° live and the maximum likelihood estimate, (also shaded darker in Figure 1b). Here, as $K \to \infty$ the estimate would approach zero. Had the maximum likelihood estimate appeared elsewhere, negative modes might not exist.

It should be noted, that the above two dimensional example is somewhat misleading. It should be remembered that the three dimensional generalizations of the priors are not symmetrical cylinders or cones but "flattened" cylinders and cones. That is, the isolikehood cylinder in the three dimensional case with d=0 has an intersection with the plane

$\beta_1 + \beta_2 + \beta_3$ = constant which is an ellipse, not a circle. In three dimensions the curve of tangencies for normal smoothness priors with d=0 will not, in general, be a straight line.

One problem which might arise in estimation is that even in the positive quadrant the modal estimate may not be unique. If the X matrix is of the right form, we might find a dependency among the equations defining $\hat{\beta}_c$ (expression (11)) so that a unique maximum cannot be found. This situation would be analogous to that in which the Almon procedure fails when A'X'XA cannot be inverted. In addition, it is also possible under some circumstances that the posterior distribution may have more than one local maximum. It might be a good idea to search for all local maxima in order to find the global maximum. However, even in cases where multiple modes may arise, we are likely to be satisfied with the mode which lies closest to the prior mean.

IV. An Iterative Procedure

A Gauss-Newton type iterative procedure to find the constrained estimate $\hat{\beta}_c$ was chosen which is easily implemented once we have a computer program which implements the unconstrained estimate $\hat{\beta}_u$. Though our estimate will be the mode of b, it was thought convenient to deal directly with β. That is, our estimate $\hat{\beta}_c$ will maximize the expression:

$$(Y-X\beta)'(Y-X\beta) + K^2 \log(\beta)'R'R \log(\beta) \tag{12}$$

with respect to β. To do this, we will approximate the second term in the expression by a quadratic function of β in the vicinity of a guess $\bar{\beta}^{(i)}$. If we chose this approximation to be a positive definite quadratic form in $(\beta-c)$ where c is a constant, then the minimum of the approximation to the expression (12) can be found by an ordinary least squares regression involving matrices augmented by dummy observations. Along Gauss-Newton lines, we substitute the first degree Taylor approximation

$$\log(\beta) \cong \log(\bar{\beta}^{(i)}) + \mathrm{diag}(\bar{\beta}^{(i)})^{-1}(\beta-\bar{\beta}^{(i)}) \tag{13}$$

into (12):

$$(Y-X\beta)'(Y-X\beta)+K^2[\log\bar{\beta}^{(i)}+\mathrm{diag}(\bar{\beta}^{(i)})^{-1}(\beta-\bar{\beta}^{(i)})]'R'R[\log\bar{\beta}^{(i)}+\mathrm{diag}(\bar{\beta}^{(i)})^{-1}(\beta-\bar{\beta}^{(i)})] \tag{14}$$

The maximum $\bar{\beta}^{(i+1)}$ of (14) is then given by:

$$\bar{\beta}^{(i+1)}=[X'X+K^2\mathrm{diag}(\bar{\beta}^{(i)})^{-1}R'R\mathrm{diag}(\bar{\beta}^{(i)})^{-1}]^{-1}[X'Y-K^2\mathrm{diag}(\bar{\beta}^{(i)})^{-1}R'R\log\bar{\beta}^{(i)}]$$

$$=[\tilde{X}'\tilde{X}]^{-1}\tilde{X}'\tilde{Y} \tag{15}$$

where:

$$\tilde{X} = \begin{bmatrix} X \\ KR(\mathrm{diag}\bar{\beta}^{(i)})^{-1} \end{bmatrix} \qquad \tilde{Y} = \begin{bmatrix} Y \\ -KR\log(\bar{\beta}^{(i)}) \end{bmatrix} \tag{16}$$

If one already has a program which implements simple smoothness priors, it is a very easy matter to have the program build these matrices as well. The procedure, then, will be to form an initial guess $\bar{\beta}^{(0)}$ for the posterior mode, and then form $\tilde{X}$ and $\tilde{Y}$ based on this guess, and regress $\tilde{Y}$ on $\tilde{X}$ to get a revised guess $\bar{\beta}^{(1)}$. $\bar{\beta}^{(1)}$ is then used to form new matrices $\tilde{X}$ and $\tilde{Y}$ to yield a new estimate $\bar{\beta}^{(2)}$. The process is repeated until $\bar{\beta}_{i+1} = \bar{\beta}_i$ up to some tolerance. When this occurs, expression (11) is satisfied by $\bar{\beta}_{i+1}$ and $\bar{\beta}_{i+1}$ is the posterior mode $\hat{\beta}_c$.

In each iteration we can say that we are approximating the prior distribution of β by a normal distribution, that is, the Taylor representation of $R\log\beta$: $R(\log)\bar{\beta}^{(i)} + \mathrm{diag}(\bar{\beta}^{(i)})^{-1}(\beta - \bar{\beta}^{(i)})$ rather than $R\log\beta$ itself is assumed spherically normally distributed with zero mean and with variance ξ.

In terms of the isodensity contours displayed in figure 1, we see that the approximation substitutes a system of parallel lines for the system of intersecting lines in 1b in such a way that the isodensity contour of the approximation which passes through the guess $\bar{\beta}^{(i)}$ coincides with the actual isodensity contour in 1b which passes through this point.

If the regression program prints standard errors of the coefficients, then these too will have an interpretation in

terms of the approximating normal priors. Under the assumption that the estimated standard error of the regression is the true standard error, then the standard error of the coefficients printed by the program will be the posterior standard errors based on the approximating normal prior. If the standard error of the regression σ is not known but is given a prior distribution $f(\sigma) \propto 1/\sigma$ then, under the assumption that the ratio $K = \sigma/\xi$ is known, the marginal posterior of each coefficient will have a student distribution with scale parameter equal to the standard error printed and degrees of freedom as printed by the program (see Zellner [1971]).

The iterative procedure may be compared with the Newton-Rapheson Method. The complete Newton step for maximizing (12) would be:

$$\bar{\beta}^{(i+1)} = [\, X'X + K^2 \mathrm{diag}(\bar{\beta}^{(i)})^{-1} R'R \mathrm{diag}(\bar{\beta}^{(i)})^{-1} - K^2 \mathrm{diag}(\bar{\beta}^{(i)})^{-2} \mathrm{diag}(R'R \log \bar{\beta}^{(i)}) \,]^{-1} \times [X'Y - 2K^2 \mathrm{diag}(\bar{\beta}^{(i)})^{-1} R'R \log \bar{\beta}^{(i)}] \qquad (17)$$

which cannot in general be implemented by a regression technique involving dummy observations, so that the procedure is less convenient from our point of view. However, we note that if we choose our initial guess $\bar{\beta}^{(0)}$ so that $R \log \bar{\beta}^{(0)} = 0$ (i.e. so that the guess is itself a truncated Koyck, normal density etc.) then the two procedures will be identical for the first iteration. If in subsequent iterations $R \log \beta$ remains small, subsequent iterations will also be similar, and our iterative procedure will show approximately quadratic convergence.

V. An Illustrative Example

To illustrate the application of the estimator based on log smoothness priors, we have chosen an example in which simple smoothness priors do not perform as well as we'd like. This is a case in which we expect the coefficients to be positive and yet the final estimated coefficients (estimated without endpoint constraints) do not "tail off" at the end but instead become negative. The equation estimated is a term structure equation developed originally by Modigliani and Sutch which relates long term interest rates to a distributed lag on past short term rates of interest. Modigliani and Sutch hypothesized that long term interest rates are determined by expectations of future short rates of interest which in turn are related to a long distributed lag on past short rates of interest. The distributed lag, they asserted, should be smooth except that the first coefficient in the distributed lag (i.e. that corresponding to the contemporaneous short rate of interest) might differ substantially from the others due to an impact effect of the current short rate. They estimated the relation with the Almon polynomial constraint that did not constrain the first coefficient of the lag. The relation was improved and reestimated using the estimator based on first degree smoothness priors in Modigliani and Shiller [1973], and was also discussed in Shiller [1973]. It was discovered at this time that if the distributed lag is extended to 24 quarters, that the "tail" of the distributed lag becomes

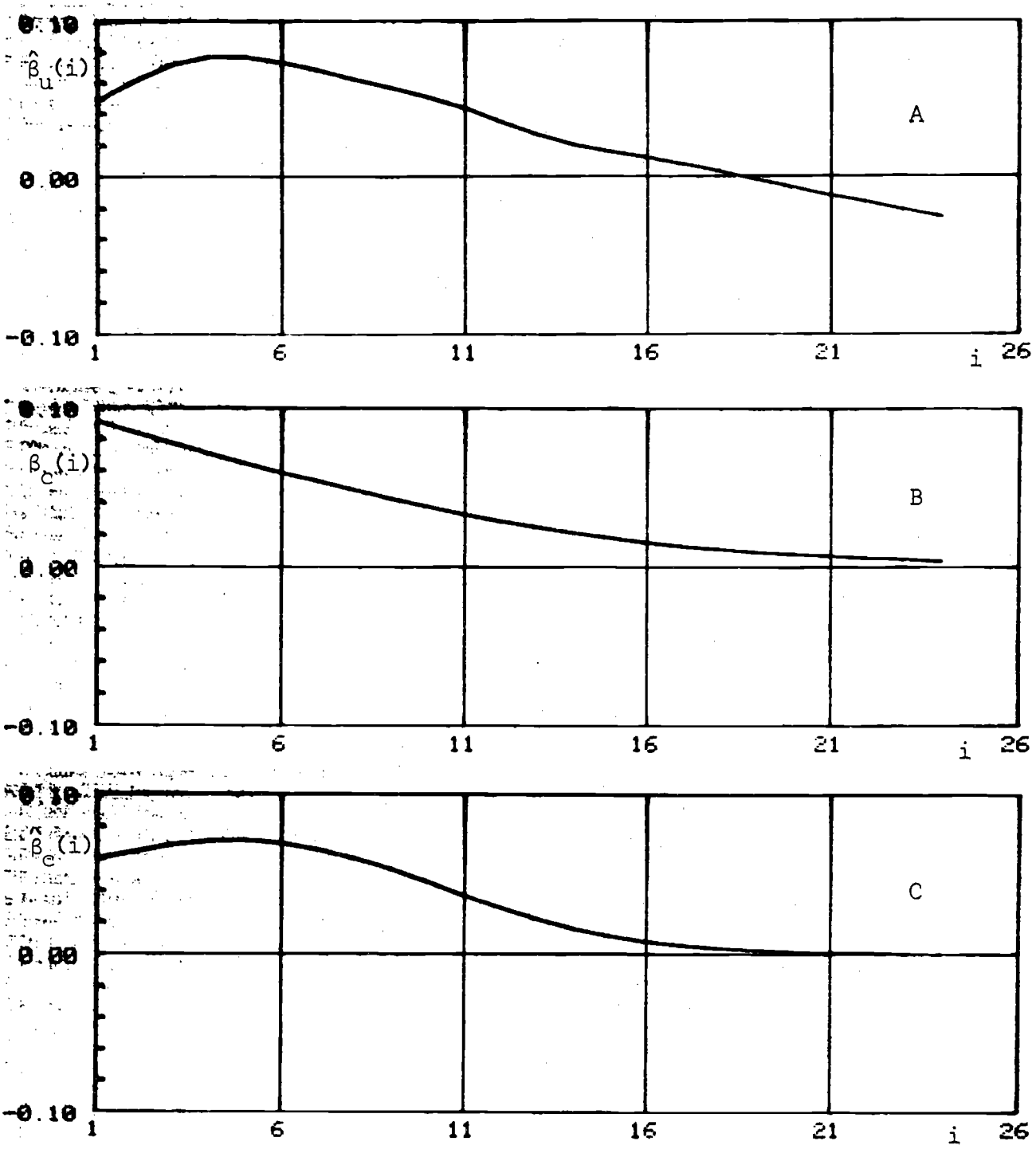

Figure 2. Estimates based on smoothness priors of β_i. A. (upper plot) unconstrained sign estimate, d=1, k=20, sum of coefficients is 1.10, R^2=.978. B, (center plot) constrained sign estimate, tight priors, d=1, k=20, sum=1.12, R^2=.974. C, (lower plot) constrained sign estimate, loose priors, d=1, k=3, sum=1.12, R^2=.977. Coefficient sums include β_0 (not plotted). R^2 is computed from original data only, excluding dummy observations.

negative. Prior information that coefficients should not be negative was then incorporated in an ad hoc manner by merely truncating the lag after 16 quarters. It would be interesting to see, then, what kind of lag curve our log smoothness prior estimate will produce when the lag is extended over the entire 24 quarter period and estimated without endpoint priors.[4]

The unconstrained sign estimate $\hat{\beta}_u$ of the distributed lag with K = 20 and d = 1 appears (except for the first coefficient) in figure 2a.[5] The shape is roughly as expected except for the negative coefficients in the tail. If the curve is reestimated using endpoint priors, then the lag curve is still negative in the tail but heads up at the end.

An Almon procedure using a third degree polynomial constraint produces an estimate which is similar to that in figure 2a in that the final coefficients are negative. An Almon estimate using a third degree polynomial constraint with the far end point tied to zero also produced negative coefficients in the tail. Among these procedures, the best way to impose non-negativity in the tail appears to be to truncate the lag distribution, i.e. choose a shorter lag length λ.

Ordinary least squares produced the typical jagged erratic estimate of the lag coefficients. The coefficients frequently alternated in sign, and ranged from -.15 to .24. No semblance of the above estimates could be seen in the ordinary least square estimate.

Using the absolute value of the estimate in figure 2a as an initial guess, the iterative procedure described above was used to find the constrained sign estimate $\hat{\beta}_c$ with d = 1 and with K = 20. In each iteration the absolute value of the coefficient estimated in the previous iteration was used as the guess $\bar{\beta}$. The first two iterations still produced negative values in the tail, but all negative values disappeared by the third iteration. The coefficients had all converged to two decimal place accuracy by the sixth iteration. The estimated lag curve, which appears in figure 2b, looks like a truncated geometric (Koyck) lag. This is not surprising, since a value of K = 20 represents a tight prior. Based on the standard error estimated at .28 in the first simple smoothness prior estimates, this prior variance on the second differences of the logs of the coefficients is only $\frac{.28}{K}$ = .014. That is, the priors assert that the ratio of a pair of adjacent coefficients should probably not differ from the ratio of the next pair of coefficients by more than 1 or 2%. In our estimate, $\hat{V} = R\log\hat{\beta}$ has its largest (in absolute value) element just after the hump, where it reaches -.008. All elements of $\hat{V}$ are negative, so the estimated curve does deviate systematically from a geometric lag. althrough not to a great degree.

In order to view the effects of a weaker log smoothness prior, K was changed from 20 to 3 and the iteration continued, using as a starting guess $\bar{\beta}_o$ the estimate that had been reached

with K = 20, and again substituting in each iteration the absolute value of the estimate from the previous iteration as the guess $\bar{\beta}_i$. Negative values of the coefficients in the tail of the lag distribution immediately reappeared in the first iteration. Since by the second iteration the negative coefficients did not seem to be getting smaller, the tail coefficients were changed to small positive numbers in the guess $\bar{\beta}$ that was used for the third iteration. There were then no negative coefficients estimated in the third iteration, and convergence to two place accuracy was again achieved by the sixth iteration. The estimate appears in figure 2c. The estimate no longer appears as a geometric lag, but rather has a humped shape which resembles (except for the tail) the shape which appears in 2a. The priors which produced 2c are much weaker than those which produced 2b: if $\beta = .28$ then $\xi = \frac{\sigma}{K}$ is about .1; i.e. our priors assert that the ratio of a pair of adjacent coefficients should probably not differ from the ratio of the next pair of adjacent coefficient by much more than 10%. In the estimate the largest (in absolute value) element of $\hat{V} = R\log\hat{\beta}$ is -.041. Apparently the likelihood function does not carry much information on $V = R\log\beta$, so that the estimates of V will come out close to the prior mean. This means that even with very loose priors, the estimated distributed lag will appear relatively smooth. The smallest (in absolute value) elements of $\hat{V}$ were those which

correspond to the tail of the distribution, apparently because the likelihood function carries very little information about the ratios of such small coefficients.

Figure 2c illustrates well some of the best properties of the estimator with constrained sign. It should be remembered that no constraints were placed on the lag of any kind and the priors were uninformative on each coefficient considered separately. The prior did not contain information that the final coefficients were small. The asymptote which appears here is the result of the interaction of the likelihood function with the prior information that coefficients must not be negative plus the information that no single coefficient should be large in the neighborhood of other small coefficients. It should not be concluded that an asymptote occurs only when the unconstrained estimate is negative in the tail. Experience with the estimator shows that asymptotes will also arise when the final coefficients are positive. The estimator with constrained sign naturally forms asymptotes (especially when d = 1 or higher) in cases where the final coefficients are small, and thus makes the estimates much less sensitive to the error of setting λ, the lag length, too high.

VI. Conclusion

Both estimators: $\hat{\beta}_u$ with unconstrained sign and $\hat{\beta}_c$ with constrained sign, should be useful in different applications. The unconstrained sign estimator may be used in cases in which there is no theoretical presumption that all distributed lag coefficients should be positive. Since the unconstrained sign estimator has more straightforward properties and does not require an iterative procedure, it may also be the choice in cases in which there is a presumption that all lag coefficients be positive. It can also form asymptotes, although it depends on the data more to make this happen. Information that all coefficients should be positive can also be used in an informal "Bayesian" approach by estimating the coefficients for several different lag lengths and choosing a truncation point that leaves all coefficients positive.

In cases in which we know all coefficients are positive we may also wish to consider whether the log normal smoothness prior might not represent our prior information sufficiently better to warrant the greater computational burden of the constrained sign estimator. It is easier to specify the parameters of the log smoothness prior since they are expressed in percentage terms and are hence unit free. The estimator is less sensitive to an overstatement of the lag length since it easily forms asymptotes. Since the estimator essentially

embodies Leamer's principle of proportionality, isolated large coefficient estimates in the tail of the lag distribution are effectively prevented. So little damage is caused by over-estimating the lag length λ that one might use for λ one's upper bound to the possible true lag length in the estimator. Finally, the limiting behavior of the constrained sign estimator as the tightness of the prior goes to zero is probably more acceptable than is the case with the unconstrained sign estimator. The limiting constraints with the log smoothness prior are, for $d = 1$ the truncated geometric (Koyck) constraint, for $d = 2$ a truncated normal density constraint, rather than polynomial constraints as is the case with smoothness priors.

Footnotes

1 For the fundamentals of Bayesian econometrics, see Zellner [1971].

2 The priors approach the constraint the the coefficients lie on a d degree polynomial. Almon also constrained the polynomial to pass through zero at the head (β_{-1}) and tail (β_{λ}). To make the analogy to the Almon procedure complete, we can take the spherically normal prior on $u = R\beta$ where R is a $(\lambda-d-1+h+t \times \lambda)$ matrix whose i.j.th element is zero if $j-i>d+1-h$ or if $i>j+h$ and otherwise equals $(-1)^{j+h-i} \binom{d+1}{j+h-i}$ where h = 1 if the head is constrained and is zero otherwise, t = 1 if the tail is constrained and is zero otherwise. These priors then include the zero "coefficients" beyond the lag and approach the Almon constraint as $\xi \to 0$. Henceforth in this paper we refer to the Almon polynomial constraint <u>without</u> head and tail constraints.

3 Leamer formulated his "Principle of Proportionality" for fully informative multivariate normal priors on the β_i. His principle actually states that the prior standard deviation of the β_i should be inversely proportional to the prior mean of the β_i.

4 It should be noted that the result of using endpoint priors in the log smoothness priors case which are analogous to the endpoint priors in the smoothness priors case (footnote 2 above) amounts to assuming prior information that the final coefficients lie near 1 rather than zero.

5 Estimation was done with a program written by Stanley Wasserman, which took the form of a MACRO file on the TROLL system. The MACRO, which is entitled &SHILLER, is available to users of the TROLL system, but cannot be used separately from the system. The dependent variable is a version of the Federal Reserve Board new issue yield series, formed by splicing an older unpublished series to their published series which starts in 1960. The independent variable is the 4 to 6 month prime commercial paper note. The sample period is 1955 second quarter to 1974 fourth quarter.

References

Almon, Shirley, "The Distributed Lag between Capital Appropriations and Expenditures", Econometrica 33 (1965); 178-196.

Cleveland, William S., "Estimation of Parameters in Distributed Lag Econometric Models" in Zellner and Fienberg, editors, Studies in Bayesian Econometrics and Statistics, North Holland, 1975.

Griliches, Zvi, "Distributed Lags: A Survey", Econometrica 35 (1967): 16-49.

Hill, R. Carter and S.R. Johnson, "Distributed Lag Estimators Derived from Smoothness Priors: A Comment", unpublished paper, Department of Economics, University of Missouri - Columbia, 1975.

Leamer, Edward E., "A Class of Informative Priors and Distributed Lag Analysis", Econometrica 40 (1972): 1059-81.

_____, "Regression Selection Strategies and Revealed Priors", Discussion Paper, Harvard Institute for Economic Research, 1975.

_____, "Ridge Regression Metrics and Distributed Lag Coefficients", Discussion Paper, Harvard Institute of Economic Research, Cambridge 1974.

Maddala, G.S., "Ridge Estimators for Distributed Lag Models", NBER Computer Research Center Working Paper No. 69, Cambridge, Mass., 1974.

Modigliani, Franco and Robert J. Shiller, "Inflation, Rational Expectations and the Term Structure of Interest Rates", Economica, Feb. 1973: 12-43.

Shiller, Robert J., "A Distributed Lag Estimator Derived from Smoothness Priors", Econometrica 41 (1973): 775-88.

Sims, Christopher A., "Distributed Lags" in Intriligator, ed., Frontiers of Quantitative Economics Volume II, North Holland, 1974.

Wilson, John F., "Have Geometric Lag Hypotheses Outlived their Time?", unpublished manuscript, Board of Governors of the Federal Reserve System, Washington, D.C., 1975.

Zellner, Arnold, An Introduction to Bayesian Inference in Econometrics, New York, Wiley, 1971.