The Annals of Mathematical Statistics
1971, Vol. 42, No. 5, 1622–1636

DISTRIBUTED LAG ESTIMATION WHEN THE PARAMETER SPACE IS EXPLICITLY INFINITE-DIMENSIONAL

BY CHRISTOPHER A. SIMS
University of Minnesota

1. Introduction. This paper discusses a number of results which were developed for application to the model

$$y = x * b + \varepsilon, \tag{1}$$

where x and ε are independent covariance–stationary stochastic processes with zero mean, b is a square summable sequence of real numbers, and "$*$" denotes convolution.[1] We will consider only the case with discrete time parameter.[2]

Where b is known to lie in some finite-dimensional linear space of sequences, estimation of b in (1) from a sequence of observations on y with x known can be carried out by least squares or generalized least squares. Even where restriction of b to such a space can only be regarded as approximately accurate common practice is to proceed with estimation as if the model were finite–dimensional without explicit concern for the effects of approximation.

There is one trivial case in which it is obviously possible to obtain consistent estimates with consistent confidence statements in an infinite–dimensional parameter space for b. Suppose b is known to lie within

$$S = \bigcup_{j=1}^{\infty} A_j, \tag{1A}$$

where A_j is a finite-dimensional linear space containing its predecessors. Any reasonable metric on S[3] will induce a topology on A_j equivalent to Euclidean topology. A natural procedure, then, is to start with A_1, estimating b in (1) on the assumption that b in fact lies in A_1. When the estimated confidence region for b has been reduced in radius (in the relevant metric) to $\frac{1}{2}$, proceed to A_2. Continue in this manner, shifting to A_{j+1} in every case only when the confidence region estimated within A_j has shrunk to a maximum radius of 2^{-j}. Even though the estimates and confidence regions may be inaccurate to start with, they are bound eventually to become accurate when we reach a j large enough so that the true value of b (call it b_0) lies in A_j. The only question is whether we can be sure that the procedure will make j go to infinity with probability one. This condition is easily verified for, say, the case where x and ε both have spectral densities bounded away from zero and infinity and are ergodic.

Received April 17, 1970; revised February 16, 1971.

[1] I.e., $a*b(t) = \sum_{s=-\infty}^{\infty} a(s)b(t-s) = \sum_{s=-\infty}^{\infty} b(s)a(t-s)$.

[2] The effects of approximating a model of the form (1) but with continuous time parameter by a similar model with discrete time parameter have been considered by the author (1971). A discussion of the implications of this paper's approach and results for econometric practice appear in Sims (1969).

[3] Any which makes it a topological vector space.

But these results are not of much practical importance. In my view, asymptotic results in statistics are useful primarily as convenient paradigms to describe what will happen as we proceed from small to very large samples. In many or most practical applications of (1), we can pass from small to very much larger samples without ever reaching a situation where it is practical to introduce a parameter space which we know with certainty to contain the true b. We may nonetheless be able to introduce parameter spaces which we are confident contain b's very close to the true b in some sense.[4] In quarterly econometric models for example, we might be quite sure that the true $b(t)$ vanishes for $t > 80$, and also quite sure that we can choose five- or ten-dimensional linear spaces which already contain very good approximations to the true b. If we want to regard sample sizes of 100 or so as "large," a theory which assumes that no finite-dimensional parameter space contains the true b may be more useful to us than one which exploits only the fact that b vanishes for $t > 80$.

For these reasons, I think the natural choice for a parameter space in the estimation of (1) with approximate restrictions is an infinite-dimensional linear space of real-valued sequences with a *complete* metric topology. A space S like that defined in (1A) cannot be complete in a metric topology if the A_j's are of strictly increasing dimension. Furthermore, if the space S defined in (1A) is regarded as a subset of its completion, S is small in the topological sense that it is a countable union of nowhere-dense sets, i. e. meagre.[5]

Since a number of the theorems to follow depend for their impact on the notion that a meagre set is "small" and its complement "large," it is worthwhile examining for a moment the justification for this identification. The only solid justification is that the complement of a meagre subset of a complete linear space is always dense in the space and itself non-meagre. But of course meagre sets may themselves be dense, and on the unit interval there are meagre sets with an independent claim, arising out of the algebraic and topological structure of the interval, to being "large": there are meagre subsets of the unit interval with Lebesgue measure one.[6] In an infinite-dimensional linear space with a complete locally convex metric topology there is no translation-invariant σ-finite Borel measure—i.e., no analogue to Lebesgue measure[7]—and hence no natural alternative to "meagreness" as a definition of "smallness." Still, it should be borne in mind that a meagre set is small only in a special sense.

---

[4] See Jorgenson (1966) for an earlier discussion of approximation in models like (1). In econometrics, such models are called "distributed lag" models and the functions b "lag distributions." Jorgenson showed that certain unions of finite-parameter families of lag distributions were dense in the space of all real sequences in the uniform metric. He failed to note, though, the importance of the choice of metric to approximation questions in infinite-dimensional space and did not consider explicitly the effects of approximation error on inference.

[5] A finite-dimensional subspace of a complete, metric, infinite-dimensional linear space is nowhere dense. See Schaefer (1966) page 21 ff.

[6] See, e.g., Munroe (1953) page 113 f.

[7] See Gel'fand and Vilenkin (1964) page 359 ff.

Two approaches to estimation in infinite-dimensional spaces are taken up in this paper. First, we look at the usual practice of using approximate finiteness restrictions and ask what are the conditions under which, by slowly relaxing restrictions as sample size increases, we can make approximation error in some sense asymptotically negligible. Second, we look at an explicitly Bayesian approach, asking conditions for its consistency and comparing the "restrictions" implicit in the placing of a prior measure on an infinite-dimensional space with those required to justify the successive approximations approach.

2. Successive approximations. Suppose S is some subset of the space R^∞ of one-sided sequences $\{b(t)\}_{t=0}^{\infty}$. Define an inner-product on S via the norm

$$g_x(b) = (E[(x * b)^2])^{\frac{1}{2}}.$$

It is not hard to verify that the b in a subset A of S which minimizes $E[(y-x * b)^2]$ when the true value of b in (1) is b_0 will be the g_x-projection on A of b_0. Suppose that we can form a sequence $\{A_m\}$ of subsets of S within each of which we can form a consistent estimate of b_m, the g_x-projection on A_m of b_0. Suppose further that the A_m can be chosen so as to have g_x-dense union in S. One might hope in this situation that by proceeding slowly down the A_m sequence as sample size increases, one could obtain g_x-consistent estimates of b_0. One might even hope to use the estimates within A_m's to form confidence regions for b_0.

These hopes will not be justified unless we can find some rule for choosing a sequence $m(n)$ such that, if $\hat{b}_{mn}$ is the estimate of b_m we form the nth sample, $g_x(\hat{b}_{m(n)n}-b_{m(n)})$ converges in probability to zero. Furthermore, if the estimates within A_m's are to be used to form confidence regions for b_0, we must be able to determine some properties of an asymptotic distribution for $d(\hat{b}_{m(n)n}, b_{m(n)})$ under a relevant metric d.

If this sort of approach is to apply in practice, the relation of the g_x-topology to natural loss functions on S is important. I have discussed this point in (1969), where it is shown that g_x induces the same topology as the mean square norm g_2,

$$g_2(b) = (\textstyle\sum_{t=0}^{\infty} b(t)^2)^{\frac{1}{2}}, \tag{2}$$

so long as x has an everywhere positive bounded spectral density.

Let A_m be the space spanned by the first m elements of the sequence $\{c_j\}_{j=1}^{\infty} \subset S$. Let $\mathbf{Z}_{mn}$ be the $n \times m$ matrix with typical element $c_j * x(i)$, y_n the $n \times 1$ vector with typical element $y(i)$, ε_n the $n \times 1$ vector with typical element $\varepsilon(i)$, and so on. The least squares estimate of b_m is $\hat{b}_{mn} = \sum_{j=1}^{m} \hat{a}_{mn}(j)c_j$, where the vector $\hat{a}_m$ is defined by

$$\hat{a}_{mn} = (\mathbf{Z}'_{mn}\mathbf{Z}_{mn})^{-1}\mathbf{Z}'_{mn}y_n.$$

The vectors $\hat{a}_{mn}$ may be thought of as estimates of the underlying vector a_{m0}, defined by the relation

$$b_m = \textstyle\sum_{j=1}^{m} a_{m0}(j)c_j.$$

Define $\lambda_{\max}(k,j)$ and $\lambda_{\min}(k,j)$ as the maximum and minimum eigenvalues (respectively) of $\mathbf{Z}_{kj}'\mathbf{Z}_{kj}$. Then we can show:

THEOREM 1. *Suppose*: (i) $m(n)$ *is a sequence of integers diverging to infinity such that with probability one* $\liminf(\lambda_{\min}(m(n),n)/\lambda_{\max}(m(n),n)) > 0$; (ii) $\bigcup_{j=1}^{\infty} A_m$ *is* g_x*-dense in* S; (iii) ε *has a spectral density bounded away from zero. Then*

$$(3)\qquad E[(\hat{a}_m - a_{m0})'(\hat{a}_m - a_{m0}) \mid x]/E[\varepsilon'\mathbf{Z}_m(\mathbf{Z}_m'\mathbf{Z}_m)^{-1}(\mathbf{Z}_m'\mathbf{Z}_m)^{-1}\mathbf{Z}_m'\varepsilon \mid x]\ ^{8}$$

converges in probability to one as $n \to \infty$.

PROOF. Set $\eta_m = y - x * b_m - \varepsilon = (b_0 - b_m) * x$. It is then easily verified that we can rewrite (3) as

$$(4)\qquad 1 + \frac{E[\eta_m'\mathbf{Z}_m(\mathbf{Z}_m'\mathbf{Z}_m)^{-1}(\mathbf{Z}_m'\mathbf{Z}_m)^{-1}\mathbf{Z}_m'\eta_m \mid x]}{E[\varepsilon'\mathbf{Z}_m(\mathbf{Z}_m'\mathbf{Z}_m)^{-1}(\mathbf{Z}_m'\mathbf{Z}_m)^{-1}\mathbf{Z}_m'\varepsilon \mid x]}.$$

Define

$$\mathbf{C}_m = \mathbf{Z}_m(\mathbf{Z}_m'\mathbf{Z}_m)^{-1}(\mathbf{Z}_m'\mathbf{Z}_m)^{-1}\mathbf{Z}_m'.$$

Then (4) can be rewritten

$$(5)\qquad 1 + \operatorname{tr}(\mathbf{C}_m\eta_m\eta_m')/\operatorname{tr}(\mathbf{C}_m E[\varepsilon\varepsilon']),$$

where "tr" is the trace operator. In obtaining (5) we use the facts that ε is independent of x and that $\mathbf{Z}_m$ and η_m are both determined by x. We now take note of a well-known algebraic inequality:

$$(6)\qquad \lambda_{A\max}\operatorname{tr}(B) > \operatorname{tr}(AB) > \lambda_{A\min}\operatorname{tr}(B),$$

where $\lambda_{A\max}$ and $\lambda_{A\min}$ are, respectively, maximum and minimum characteristic roots of $\mathbf{A}$. It is easily shown that $\operatorname{tr}(\mathbf{C})_m$ exceeds $n/\lambda_{\max}$. From the discussion in Grenander and Szego (1958) pages 63–64, the minimum characteristic root of $E[\varepsilon\varepsilon']$ is known always to exceed $S_{\varepsilon\min}$, the minimum of the spectral density function for ε. It can be shown that the non-zero characteristic roots of $\mathbf{C}$ are those of $(\mathbf{Z}'\mathbf{Z})^{-1}$, so that the maximum characteristic root of $\mathbf{C}$ is less than $1/\lambda_{\min}$. Finally,

$$(7)\qquad (1/n)\operatorname{tr}(\eta_m\eta_m') = (1/n)\textstyle\sum_{t=1}^{n}\eta_m(t)^2,$$

so that

$$(8)\quad (1/n)E[\operatorname{tr}(\eta_m\eta_m')] = \operatorname{Var}(\eta_m(t)) \to 0 \quad \text{as} \quad m \to \infty \qquad \text{(by assumption (ii))}.$$

But a sequence of positive random variables whose expectations converge to zero converges in probability to zero. Thus

$$(9)\qquad \operatorname{plim}_{n\to\infty}(1/n)\operatorname{tr}(\eta_m\eta_m') = 0.$$

[8] In expression (3), the argument of the $m(n)$ functions has been suppressed for legibility. All "m's" in (3) are to be thought of as dependent on n. This same convention will be followed through most of the proof below, where n would appear as subscript or argument.

Using (6) through (9) above we can take the probability limit of (5) to get

$$\operatorname{plim}\left[1+\operatorname{tr}(\mathbf{C}_m\eta_m\eta_m')/\operatorname{tr}(\mathbf{C}_m E[\varepsilon\varepsilon'])\right]$$
$$= 1+\operatorname{plim}\left[(\lambda_{\max}/(\lambda_{\min} S_{\varepsilon\,\min}))(1/n)\operatorname{tr}(\eta_m\eta_m')\right] = 1. \quad \square$$

The hypotheses of Theorem 1 are rather general and abstract. To see that they apply to some useful cases, consider what we can say when the c_j's are all right-translates of c_1—that is, $c_j(t) = c_1(t-j)$. A sufficient condition for Theorem 1's assumption (ii) to hold can be developed from the following lemmas:

LEMMA 1. *If $\tilde{q}(\omega)$ is* (i) *defined in the lower half of the complex plane* (*including the real line*), (ii) *bounded above in absolute value*, (iii) *continuous over its domain, and* (iv) *periodic with period 2π as a function of its real part for any fixed value of its imaginary part, then $\tilde{q}$ is the Fourier transform of a square-summable lag distribution q on the integers for which $q(t) = 0$ for $t < 0$.*

PROOF. Here we are thinking of q as a sequence of discrete weights at the integers along the real line, which define a "generalized function" or "Schwarz distribution." From this point of view, q's Fourier transform is a periodic function, and in fact, any function on the real line with period 2π which is square-integrable over finite intervals is the Fourier transform of a square-summable set of weights on the integers (see Lighthill (1964)). What we need to verify here is only that q is "right-handed," i.e., that $q(t) = 0$ for $t < 0$. For this we draw on the Paley–Weiner Theorem (see Yosida (1965), page 163). That theorem shows that any function $f(\omega)$ which is analytic in the lower half plane and satisfies

$$\sup_{y<0}\int_{-\infty}^{\infty} f(x+iy)\,dy < \infty$$

is the Fourier transform of a square-integrable function on the real line which vanishes for negative arguments. Set $a(t) = 1$ for $0 \leqq t < 1$, $a(t) = 0$ elsewhere. Then $q(t)$, the square-summable sequence of which $\tilde{q}$ is the Fourier transform, is right-handed if and only if $q * a$ is a square-integrable function on the real line which vanishes for negative arguments. Thus this lemma is proved if $\tilde{q}\tilde{a}$ satisfies the hypotheses of the Paley–Weiner theorem. But $\tilde{a}$ itself satisfies those hypotheses, and the assumptions of this lemma then guarantee that $\tilde{q}\tilde{a}$ does as well.

LEMMA 2. *Linear combinations of finite numbers of right-translates of a* (*discrete*) *lag distribution c are g_x-dense in the space of all lag distributions for which g_x is finte if* (i) *x has a spectral density bounded away from zero and infinity and* (ii) *the Fourier transform $\tilde{c}$ of c satisfies $|\tilde{c}(\omega)| > \varepsilon > 0$ for all ω with non-positive imaginary part.*

PROOF. If $\tilde{b}/\tilde{c}$ is the Fourier transform of a square-summable right-handed distribution q, then we can write $b = c * q$. Under assumption (i) of this lemma, g_x induces the g_2-topology (see page 1624). Absolutely summable b's are g_2-dense in the space of square-summable distributions. Hence, if for every absolutely summable b we have $\tilde{q} = \tilde{b}/\tilde{c}$ for a right-handed, square-summable q, finite linear combinations of right-translates of c are g_x-dense. But assumption (ii) of this lemma is easily seen to guarantee that $\tilde{q} = \tilde{b}/\tilde{c}$ satisfies the assumptions of Lemma 1 for any absolutely summable b.[9] □

Lemma 2 shows that assumption (ii) of Theorem 1 is satisfied if $|\tilde{c}_1|$ is bounded away from zero and infinity in the lower half-plane. There remains assumption (i). Rules for choosing $m(n)$ so that assumption (i) of Theorem 1 is satisfied can be formulated under suitable restrictions on x and c_j. If an infinite past for x is known and the c_j's are translates of a fixed right-handed lag distribution, choosing $m(n)$ can be based naturally on knowledge of the upper and lower bounds of the spectral density of $c_1 * x(t)$. But of course in practice neither an infinite past for x nor any other source of bounds on S_x are usually known, even when we might be willing to assume that S_x is bounded away from zero and infinity. In this case in order to guarantee assumption (i) we require some way of bounding the dispersion of $\lambda_{\max}/\lambda_{\min}$ which is consistent uniformly in m as n goes to infinity. Such a bound is possible, though to give an explicit rule for determining $m(n)$ here would take us too far afield. Clearly, we could form a uniformly consistent bound on the dispersion of $\lambda_{\max}/\lambda_{\min}$ if we could form a uniformly consistent bound on the dispersion of the individual elements of $(1/n)\,Z'Z$. In the case at hand, with the c_j's all translates of c_1, $(1/n)Z'Z$ is the sample covariance matrix of the first m lagged values of a covariance-stationary process. Hence uniformly consistent bounds on the variances of the elements of $(1/m)Z'Z$ are available from the formulas given on page 39 of Hannan (1960) for, e.g., the case where x is Gaussian and $c_1 * x$ has square-summable spectral density.

Theorem 1 gives conditions under which inference carried out under the incorrect assumption that b_0 lies within the current approximating space will eventually yield accurate location estimates and accurate Chebyshev-type confidence bounds for the d_x-projection on the current approximating space of b_0. The question then arises whether we can use estimates of d_x-projections on the A_i to estimate b_0 within the parameter spaces larger than A_i. In one sense, the answer is certainly yes. If S_0 is a set which can be *uniformly* approximated with arbitrarily high accuracy by elements of the sequence A_i, and if the relevant metric is weaker than d_x, we can obviously combine consistent confidence statements about the location of b_m within A_m with knowledge of how well A_m approximates S_0 to obtain consistent confidence statements about b_0. However, the sets S_0 which can be uniformly approximated by finite-dimensional A_i turn out to be small in a topological sense.

THEOREM 2. *If S_0 is a subset of a linear space S with a metric topology defined by the metric d, and for any $\delta > 0$ there is a finite-dimensional linear subspace A of S such that for every b in S_0 there is a b' in A satisfying $d(b, b') < \delta$, then S_0 is locally precompact.*

---

[9] Here and elsewhere we will be using the fact that the Fourier transform of an absolutely summable, right-handed sequence is bounded and continuous along the real line and in the lower half-plane.

PROOF. Pick an arbitrary sphere U in S with radius, say, r, and pick an arbitrary $\delta > 0$. By hypothesis we can pick a finite-dimensional subspace A of S which approximates S_0 to within $\delta/2$. Consider the set U^* of all points in A which lie within $\delta/2$ of the set $S_0 \cap U$. U^* is clearly contained within a sphere of radius $r+(\delta/)2$ within A, and by the local compactness of Euclidean space, can therefore be covered by a finite number of spheres of radius $\delta/2$. Form a new set U^{**} by taking the spheres in the finite covering of U^*, doubling their radius, and setting U^{**} equal to their union. U^{**} will then cover $S_0 \cap U$. But since δ was arbitrary, we have proved precompactness for $S_0 \cap U$. □

Theorem 2 sharply restricts the nature of S_0. If $S_0 = S$, S can be a complete linear space only if it is finite-dimensional. If S is a complete linear space, S_0 must be a nowhere-dense subset of it.[10] In fact, these restrictions can be somewhat relaxed, because we are concerned only with *asymptotically* accurate confidence statements. For this purpose it is enough if we can find a mapping f from the positive real numbers to the class of all finite-dimensional subspaces of S such that for every point b in S_0 there is a $\delta_b > 0$ such that for all $\delta < \delta_b$ there is a point b' in $f(\delta)$ such that $d(b, b') < \delta$. Suppose S_0 satisfies this condition. Then clearly

$$S_0 = \bigcup_{j=1}^{\infty} S_{0j},$$

where S_{0j} is the set of all points b in S_0 such that for all $\delta < 2^{-j}$ there is a point b' in $f(\delta)$ such that $d(b, b') < \delta$. Each of these sets S_{0j} satisfies the hypothesis of Theorem 2, so that S_0 now, while not necessarily nowhere-dense in a complete S, is certainly a countable union of nowhere-dense sets—i.e., meagre in S.

The result of Theorem 2 can be extended directly to a somewhat broader class of approximating sets A. The essential property of A for Theorem 2 was that its intersection with any g_x-sphere was precompact. On this criterion not only finite-dimensional linear subspaces of S qualify, but also any set A which is homeomorphic to a closed subset of a Euclidean space under a transformation which takes d-spheres into bounded sets. The finite-dimensional parameter spaces of rational lag distributions (see Jorgenson (1966)) form a class of A's with this property when the d-metric is the g_x-norm.

3. A Bayesian approach. The successive approximation of the parameter space by finite-dimensional spaces of increasing dimension is unsatisfactory in that, even asymptotically, it allows us to make accurate confidence statements only if b is restricted to a meagre subset of an infinite-dimensional complete space. Might we avoid these restrictions by spreading a "smooth" and fairly "flat" prior measure over the infinite-dimensional parameter space, then using sample information to modify the prior according to Bayes' formula?

Consider a simple practical example of a prior measure on the space of b's. The measure, call it μ, could put weight 2^{-n} on each of the n-dimensional subspaces E_n

[10] These assertions follow from elementary theorems about topological vector spaces. See, e.g., Schaefer (1966) page 21 ff.

of finite b's of length n. E_n is spanned by the first n unit-distributions e_i, with $e_i(i) = 1$, $e_i(t) = 0$ for $t \neq i$. Within E_n, the measure could have, say, the form of a $N(0, \sigma^2 I)$ distribution. With this prior, it is easy to verify that any linear functional of b will have a posterior expectation which is a weighted average of the posterior expectations computed separately within the E's, the weights depending on (if ε is a normal white noise) the average residual sum of squares within the various E_n's. The Bayesian procedure would, like the successive approximations method, move slowly to higher-dimensional E_n's as lower-dimensional ones proved to give less good fits. The difference would appear to be mainly in the Bayesian procedure's averaging in of results over several heavily weighted E_n's instead of looking at only one E_n at a time. (With more careful attention to the specification of the prior, a Bayes procedure along these lines would be better in small samples than the usual econometric procedure of picking a "length of lag" by some *ad hoc* method, then proceeding with inference conditioned on the choice of lag length.)

The measure does satisfy a fundamental smoothness requirement: it puts positive probability on every g_2-sphere of positive radius. Yet it obviously fails to get us out of the problem of needing to restrict b to lie in a small subspace of a complete space in order to make confidence statements. With μ as prior we have put probability zero on all truly infinite b's. Any set with μ-probability one is dense in (say) l_1 or l_2,[11] but the set of all finite lag distributions is meagre in either of those two spaces (and in fact in any complete infinite-dimensional metric space).

This situation is not the fault of μ's being too simple a measure. One instructive way to convince oneself of this is to try to construct measures on l_1 or l_2 which do not concentrate on meagre subsets. To see the same thing deductively, note that it is shown in Parthasarathy (1967), page 29, that any Borel measure on a complete, separable metric space is "tight" and that a tight measure by definition puts probability one on a countable union of compact sets. Since compact subsets of complete, separable, infinite-dimensional linear metric spaces are nowhere dense, any Borel probability measure on such a space puts probability one on a meagre subset of the space. Like confidence statements based on the successive approximation approach, Bayesian confidence statements must rest on a priori restrictions which limit b to a meagre subspace of natural infinite-dimensional parameter spaces.

4. Questions of simple consistency. In most applications of (1), confidence statements will be of prime importance, yet the preceding two sections show that they will be impossible without prior restrictions which themselves cannot usually be made with complete confidence. It is an interesting question, therefore, whether inference collapses completely if these strong prior restrictions are false. We shall see in this section that in at least two complete, infinite-dimensional metric spaces, there is a successive approximations procedure which is everywhere consistent. We shall also show that under very general conditions Bayes estimates converge in

[11] The spaces l_1 and l_2 are the completions of $\cup E_n$ under, respectively, g_1 and g_2, with $g_1(b) = \sum_{t=0}^{\infty} |b(t)|$.

probability to a false value only on a subset of a meagre set. Unfortunately, a theorem giving conditions for consistency everywhere in some complete infinite-dimensional space for a Bayes estimate is not presented. The work of Schwarz (1965) showed that such theorems are possible for the case of independent and identically distributed observations, which suggests, but is little help in proving, that they might also be possible in the more complicated model (1).

First, we recast Schwarz's Theorem 3.5 from (1965) giving general conditions for prior-probability-one consistency of Bayes estimates.

THEOREM 3 (Schwarz, Doob). *Let S be a metric space with a Borel measure P_0 on it. (Here and henceforth we mean by "measure," "probability measure.") For each b in S, let P_b be a Borel measure on the space R^∞ of sequences of real numbers $y = (y_1, \cdots, y_n, \cdots)$ under the product topology. Let $\mathscr{F}$ be the Borel field on R^∞ and $\mathscr{F}_n$ be the sub-σ-field generated by $y_1, \cdots, y_n$. If*

(i) *for each A in $\mathscr{F}_n$, $n = 1, \cdots, \infty$, $P_b(A)$ is Borel-measurable as a function of b, and*

(ii) *there is a function f from R^∞ onto S such that $f(y) = b$ w.p.1 (P_b) and for any closed sphere $U \subset S$, $f^{-1}(U) \in \mathscr{F}$, then for any bounded P_0-measurable random variable Z on S,*

$$E[Z \mid y_1, \cdots, y_n] \rightarrow Z(b) \text{ w.p.1} (P_b) \text{ w.p.1} (P_0).^{12}$$

PROOF. The theorem stated here is a slightly less general form of that stated by Schwarz (1965) as Theorem 3.5. As is indicated in that article, the proof consists in showing that (ii) allows us to identify Z with an equivalent random variable on $S \times R^\infty$ whose value is independent of its first argument. The result then follows from an application of the Martingale Closure Theorem (Loève (1960), page 394).

The next step is to show that in a wide class of situations Theorem 3 can be applied to the model (1). The propositions to follow show that if the measure P_b puts probability one on the event that asymptotic mean square residual in (1) is minimized at b, and if x is linearly regular,[13] then Theorem 3 applies. The interesting points about what follows are: (a) that P_b need not specify an exact structure for ε—P_b may be integrated over an arbitrary prior distribution on stationary ergodic structures for ε; (b) though the residual sum of squares is naturally related to the g_2-topology on the space of b's, Theorem 4's results apply to estimation of some functions of b (e.g., the sum of the $b(t)$'s) which are g_2-discontinuous; and (c) though non-regular x processes whose spectral densities vanish over some but not all intervals "distinguish points" in the sense that the property that asymptotic mean square residual has a unique minimum at the true b is preserved for these x's, Theorem 4 does not extend to such an x-process.

---

[12] The conditional expectation in the conclusion of the theorem is defined with respect to the measure on $S \times R^\infty$ obtained by integrating P_b with respect to P_0.

[13] A linearly regular process is one which cannot be linearly forecast with arbitrary accuracy from its own past. See Rozanov (1966).

Define the function $G_T(y, b)$ as

$$G_T(y,b) = (1/T)\sum_{t=1}^{T}(y(t)-x*b(t))^2.$$

THEOREM 4. *If in the model* (1)

(i) *S is a separable metric space,*

(ii) *with x-probability one $P_{b,x}$ is defined as a Borel measure on R^∞ for all b's in a Borel set $S' \subset S$,*

(iii) *with x-probability one, $G_T(y, b) \to H(b; b')$ as $T \to \infty$* w.p.l. $[P_{b,x}]$ *where $H(b; b') = g_x(b-b')+\sigma^2$ and σ^2 is a random variable defined on $\mathcal{F}$ independent of b, and*

(iv) *the metric d on S is such that the g_x-closure of a d-sphere is its d-closure, then there is with x-probability one a function f_x: R^∞ onto S' such that $f_x(y) = b$* w.p.1 $(P_{b,x})$ *and $f_x^{-1}(U)$ is a Borel set for each (relatively) closed sphere $U \subset S'$.*

PROOF. By (i), S' has a countable dense subset V. We will henceforth treat the parameter space as S' under the relative topology. Define

$$G(y,b) = \liminf G_T(y,b) \quad \text{as} \quad T \to \infty.$$

As the lim inf of a countable collection of measurable (in y) functions, G is itself measurable in y. (See Munroe (1953), page 152.) Furthermore, by (iii) we have

$$G(y,b') = H(b;b') \quad \text{for all} \quad b' \quad \text{in} \quad V \text{ w.p.1}(P_{b,x}).$$

Let N_j be a d_x-sphere of radius 2^{-j} about 0. Set

$$A_{bj} = \{y \mid [\liminf_{c \text{ in } (b+N_j)\cap V} G(y,c)] - [\liminf_{c \text{ in } \overline{(b+N_j)}\cap V} G(y,c)] < 0\}.$$

Then define

$$Y_b = \bigcup_{j=1}^{\infty} A_{bj}.$$

Set $f_x(y) = b$ for y in Y_b, $f_x(y) = 0$ when y is in no Y_b. Clearly $f_x(y) = b$ w.p.1 $(P_{b,x})$ for all b in S'. The question remains whether $f^{-1}(U)$ is measurable for closed d-spheres U. Define

$$Y_U = \bigcap_{j=1[c \text{ in } V\cap U]}^{\infty} A_{cj}.$$

The set Y_U is plainly Borel-measurable. We shall now show that the conditions of the theorem guarantee $Y_U = f^{-1}(U)$. That $Y_U \supset f^{-1}(U)$ follows immediately from the definitions of Y_b and Y_U. For any y in Y_U, there is a sequence $\{c_{yj}\} \subset (V \cap U)$ such that y is in $A_{c_{yj}j}$ for each j. We know that $(c_{yj}+N_j) \cap (c_{yk}+N_k) \neq \varnothing$ for all j, k. (Otherwise we could obtain the contradiction g.l.b.$_{(c_{yj}+N_j)}\, G(y,b)$ both greater and less than g.l.b.$_{(c_{yk}+N_k)}\, G(y,b)$.) Hence, by the fact that the radii of the N_j decrease exponentially, $\{c_{yj}\}$ is a Cauchy sequence with a g_x-limit a_y in the g_x-closure of U. It is easy to see that y is in Y_{a_y}. Hence if U is g_x-closed, as it will be by (iv) if U is d-closed, $Y_U \subset f^{-1}(U)$. Therefore $Y_U = f^{-1}(U)$. □

We now need to see how condition (iv) of Theorem 4 relates to usual sorts of d-metrics and x-processes.

LEMMA. *If g is a semi-norm of the form*

$$g(b) = \left(\sum_{t=0}^{\infty} a(t)|b(t)|^p\right)^{1/p}, \tag{10}$$

$p \geqq 1$, $a(t) > 0$ *for all t, then*

$$U = \{b \mid g(b) \leqq 1\}$$

is closed in the topology of coordinate-wise convergence.

PROOF. It is easily verified that if b_j converges coordinate-wise to c and c is outside U, then the b_j must eventually also be outside U. □

COROLLARY. *If* (i) *x is linearly regular*, (ii) *d is any metric defined by a countable (or finite) family of semi-norms of the form given in* (10), (iii) *conditions* (i)–(iii) *of Theorem* 4 *are met with* $P_0(S') = 1$, *and* (iv) $P_{b,x}$ *is measurable in b, then Bayes estimates are consistent for Borel-measurable functions on S with* P_0*-probability one.*

PROOF. If x is linearly regular, we know from Sims (1969) that g_x defines a topology stronger than coordinate-wise convergence. Hence (i) and (ii) plus the lemma give us (iv) of Theorem 4. With (iii), this yields us the conclusions of Theorem 4, which, with (iv), yield the conclusion of Theorem 3. □

It may be worthwhile to point out here just how it comes about that these results fail to apply to x-processes which are not linearly regular. As has already been pointed out, if x has a spectral density over some interval of positive length, then *one-sided* b's are distinguished by g_x. The reason is that the Fourier transforms of distinct one-sided b's cannot be identical over any interval of positive length. This means that Theorem 4's f_x function is well-defined and satisfies the requirement that $f_x(y) = b$ w.p.1 ($P_{b,x}$). However, since x's which are not linearly regular in general produce a g_x-topology weaker than coordinatewise convergence, the condition that inverse images of closed spheres be Borel-measurable will not in general be met for these x's. Theorem 4 will apply to x's which are not linearly regular, but the measurability condition on f_x restricts its application to estimation of functions of b which depend only on $\tilde{b}$ at frequencies in intervals over which x has positive spectral density. This excludes many practically interesting functions, but it should be noted that, e.g., in the case where x has non-zero spectral density only in some interval about $\omega = 0$, the sum of coefficients in $b(t)$ may be consistently estimable[14] by Bayesian methods even though no individual coefficient of $b(t)$ can be consistently estimated.

Sets of probability one need not be large in any topological sense, and we have seen that indeed in infinite-dimensional spaces there are always sets of probability one which are small in a certain topological sense, that is, meagre. Theorem 3 does allow us to say directly that if P_0 is spread smoothly over S, i.e., if it puts positive

[14] This is a conjecture. I have not actually carried out a formal analysis of what kinds of functions on S satisfy measurability conditions for non-regular x's.

probability on every sphere, then Bayes estimates are consistent on a dense set; but dense, meagre sets like the set of all finite-length b's may be "intuitively small." The following theorem gives conditions which guarantee that Bayes estimates come arbitrarily close to the true value for arbitrarily long sequences of samples on a set of b's whose complement is meagre.

THEOREM 5. *If in Theorem 3 P_0 puts positive probability on all spheres in the complete metric space S and if* (i) *is strengthened to require that $P_b(A)$ be continuous in b for sets A in $\mathscr{F}_n$, then for bounded, continuous functions Z on S, $E[Z \mid y_1, \cdots, y_n]$ comes arbitrarily close to $Z(b)$ over arbitrarily many consecutive values of n with P_b-probability one for a residual set of b's (i.e., a set with meagre complement).*

PROOF. Z will take values in some interval I. Partition I into intervals $I_{mj}, j = 1, \cdots, m$ of equal length. Let I_{m0} be the symmetric, open interval of length $1/m$ about zero. Define

$$B_{nmj} = \{b \mid P_b\{E[Z \mid y_1, \cdots, y_n] \text{in} (I_{mj}+2I_{m0})\} > 1-2^{-m}\} \cap Z^{-1}(I_{mj}+I_{m0}).$$

$$B_{nm} = \bigcup_{j=1}^{m} B_{nmj}.$$

Clearly B_{nm} is open. Any set containing all b's for which $E[Z \mid y_1, \cdots, y_n]$ converges to $Z(b)$ w.p.1(P_b) has P_0 probability one and is therefore dense. Consider

$$B^*_{mp} = \bigcup_{n=1}^{\infty} \bigcup_{k=n}^{n+p} B_{km}.$$

For each m and p, B^*_{mp} is an open set of P_0-probability one. The set of b's for which $E[Z \mid y_1, \cdots, y_n]$ comes arbitrarily close to $Z(b)$ over arbitrarily many consecutive values of n w.p.1 (P_b) includes

$$\bigcap_{m,p=1}^{\infty} B^*_{mp},$$

which has probability one (hence is dense) and is a G-delta (i.e., is a countable intersection of open sets). But in a complete metric space every dense G-delta is residual. (See Munroe (1953), page 69.) □

COROLLARY TO THEOREM 5. *Under the conditions of Theorem 5, the set of b's for which $E[Z \mid y_1, \cdots, y_n]$ converges in P_b-probability to a number other than $Z(b)$ is a subset of a meagre set.*

PROOF. The set of b's specified in the Corollary is a subset of the complement o $\bigcap_{m,p=1}^{\infty} B^*_{mp}$. □

Theorem 5 is, like Theorem 3, not directly tied to the model (1). Theorem 5's conclusions, that Bayes estimates do not behave too very badly except on a meagre set, do not apply to some of the natural specifications of (1). For example, if the x-process is linearly regular and Gaussian, $x(t)$ will be unbounded w.p.l. Hence Theorem 5's assumptions will not apply if S is the space of all absolutely summable b's under the g_1-norm

$$g_1(b) = \sum_{t=0}^{\infty} |b(t)|,$$

because $P_{b,x}$ will be continuous in b on this S only for bounded $x(t)$. There is, however, a complete S to which the Theorem does apply: The space S_s of rapidly decreasing sequences under the natural metric topology. This is the space of b's for which each of the semi-norms

$$g_{sp}(b) = \sum_{t=0}^{\infty} t^p |b(t)|, \qquad p = 0, \cdots, \infty$$

is bounded, under the topology defined by this family of semi-norms. For b's in this space, $x*b(t)$ is continuous in b w.p.1. On this space, the continuous Z's include the sum of b's coefficients and the "mean lag" defined as $[\sum_t tb(t)]/[\sum_t b(t)]$.

From the discussion in Section 2 of this paper, the following is almost immediately clear:

THEOREM 6. *When Theorem 1 applies (e.g., when the x-process has a thrice-differentiable spectral density and the nth approximating space is the space of finite distributions of length n), we can choose estimates of b which are g_2-consistent over all of l_2.*

PROOF. The argument of Section 2 showed that it is possible to choose a function $m(n)$ such that Euclidean confidence statements about the location of b_m within the $m(n)$th approximating space made from the nth sample eventually become arbitrarily accurate. Since a necessary condition for this result is that the approximating spaces be g_2-dense in the whole space of possible b's, all we need do is choose $m^*(n) \leqq m(n)$ such that our estimate of $E[g_2(\hat{b}_{m^*(n)n}, b_m)] < 2^{-m}$ for all n. This guarantees that $g_2(\hat{b}_{m^*(n)n}, b)$ converges in probability to zero, which is the result desired.

The problem here is that, though the g_2-topology is certainly of some practical relevance, a large number of interesting functions of b—including the sum of the coefficients and the mean lag—are g_2-discontinuous. Proving consistency in topologies stronger than the g_2-topology (and hence necessarily in somewhat smaller spaces than l_2) reduces to the problem of showing that given a dense family of approximating spaces, the sequence $\{b_m\}$ of g_x-projections of b on the approximating spaces A_m converges to b in the stronger topology. If S is l_1, and the A_m's are the spaces of finite distributions of length m, then the requirement that the g_x-topology be equivalent to the g_2-topology is *not* enough to guarantee that b_m converge to b relative to g_1.[15] However, it is easy to show:

THEOREM 7. *When S is S_s, the space of strongly decreasing sequences, A_m is the space of finite distributions of length m, and the g_x-topology is the g_2-topology, then b_m converges to b in the topology of S_s.*

PROOF. Each A_m contains a point b_{m0}, the g_2-projection on A_m of b (obtained, of course, by truncating b at $t = m$). The distance $g_x(b - b_{m0})$ declines faster than any

[15] I am indebted to David Ragozin, a mathematician at the University of Washington, for developing an example which proves this statement.

power of m as $m \to \infty$. Within any A_m there is a maximum ratio ρ_{pm} of g_{sp} to g_x. This ratio ρ_{pm} increases more slowly than some power of m for each p. This, together with the fact that $g_x(b_m - b_{m0}) < 2g_x(b - b_{m0})$ implies that $g_{sp}(b_m - b_{m0}) \to 0$. But since $g_{sp}(b - b_{m0}) \to 0$ also, $g_{sp}(b - b_m) \to 0$. □

Thus it turns out that the most useful consistency result for the successive approximations method applies to the same space as Theorem 5's result for Bayesian estimates.

5. Conclusion. Though the issues raised in Section 4 are interesting as abstract problems, the more important results of this paper are, I think, the negative results in earlier sections of the paper. Estimates to which no confidence statements can be attached are unsatisfactory in practice. Yet there is no way to make estimates with attached confidence statements which applies to the whole of an interesting infinite-dimensional parameter space. The effects of approximating an infinite-dimensional parameter space by finite-dimensional spaces cannot, in other words, be made asymptotically negligible. In estimating distributed lags one should, where no better alternative is available, use Bayesian procedures which satisfy the hypotheses of Theorem 5 or use finite-dimensional approximations which can be justified as part of a procedure to which Theorem 6 or 7 applies (depending on whether g_2 or a stronger topology is appropriate to the problem). But the fact that such a procedure has been used (*a fortiori* the fact that one is using an approximating space drawn from a class of spaces dense in the underlying infinite-dimensional parameter space) cannot justify ignoring approximation error, even in "large samples." The alternatives are either to give an explicit discussion of the likely nature of approximation error along the lines suggested in Sims (1969) or to develop a thoroughly convincing a priori rationale for a particular finite-dimensional parameterization, perhaps along the lines suggested by Nerlove (1967).

APPENDIX

The relation of this paper's Theorem 5 to Freedman's work. Freedman (1963), (1965) has dealt with the asymptotic properties of Bayesian inference on an "infinite-dimensional" parameter space. The model he works with is quite different from that which motivates this paper, but Theorem 5 of this paper is general enough to apply to Freedman's context. Thus we have a weak positive result to juxtapose with Freedman's negative results: Freedman shows that in his model there are priors which are "unprejudiced" in the sense that they assign positive probability to each open set but for which the posterior probability of *every* open set has a lim sup of 1 except for a meagre set of true parameter values. Theorem 5 shows that nonetheless, the posterior distribution will be arbitrarily highly concentrated near the true parameter value for arbitrarily long sequences of samples for all but a meagre set of true parameter values. An interesting but open question is whether the persistent occasional concentration of posterior probability on false values happens "less often" for larger sample sizes.

Note also that in Freedman's model it is possible to choose the prior in such a way as to guarantee consistency. His parameter space is, with appropriate choice of metric, complete metric but not linear. With an infinite-dimensional complete linear metric parameter space as in this paper it remains unknown whether the prior can be chosen to guarantee consistency for a residual set of parameter values.

REFERENCES

FREEDMAN, D. A. (1963). On the asymptotic behavior of Bayes estimates in the discrete case. *Ann. Math. Statist.* **34** 1386–1403.

FREEDMAN, D. A. (1965). On the asymptotic behavior of Bayes estimates in the discrete case II. *Ann. Math. Statist.* **36** 454–456.

GEL'FAND, I. M. and VILENKIN, N. YA. (1964). *Generalized Functions*, **4** *Applications of Harmonic Analysis*. Academic Press, New York.

GRENANDER, U. and SZEGO, G. (1958). *Toeplitz Forms and Their Applications.* Univ. of California Press, Berkeley.

HANNAN, E. J. (1960). *Time Series Analysis.* Methuen, London.

JORGENSON, D. W. (1966). Rational distributed lag functions. *Econometrica.*

LIGHTHILL, M. J. (1964). *Introduction to Fourier Analysis and Generalized Functions.* Cambridge Univ. Press.

LOÈVE, M. (1960). *Probability Theory* (2nd ed.). Van Nostrand, Princeton.

MUNROE, M. E. (1953). *Introduction to Measure and Integration.* Addision–Wesley, Reading, Mass.

NERLOVE, M. (1967). Distributed lags and unobserved components of economic time series. *Ten Economic Essays in the Tradition of Irving Fisher.* (William H. Fellner *et al.*, eds.). Wiley, New York.

PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces.* Academic Press, New York.

ROZANOV, YU. A. (1966). *Stationary Random Processes.* Holden–Day, San Francisco.

SCHAEFER, H. H. (1966). *Topological Vector Spaces*, MacMillan, New York.

SCHWARTZ, LORRAINE (1965). On Bayes procedures. *Z. Warscheinlichkeitstheorie und Verw. Gebiete* **4** 10–26.

SIMS, C. A. (1971). Discrete approximations to continuous time distributed lags in econometrics. *Econometrica.* **39** 545–563.

SIMS, C. A. (1969). The role of prior restrictions in distributed lag estimation. Harvard Institute of Economic Research Discussion Paper No. 60.

YOSIDA, K. (1965). *Functional Analysis*, 2nd printing corrected. Springer–Verlag, Berlin.