The Annals of Statistics
1988, Vol. 16, No. 2, 648–660

A SIEVE ESTIMATOR FOR THE COVARIANCE OF A GAUSSIAN PROCESS[1]

BY JAY H. BEDER

University of Wisconsin at Milwaukee

Maximum likelihood estimation for the covariance R of a zero-mean Gaussian process is considered, with no assumptions on the covariance or the "time" parameter set T. It is shown that the likelihood function is a.s. unbounded in general, and a sieve estimator $\hat{R}$ is constructed. The distribution of $\hat{R}$, considered as a process on $T \times T$, can be described exactly if a certain technical assumption is satisfied concerning the bivariate series expansion of R. It is then shown that $\hat{R}(s, t)$ is asymptotically unbiased and consistent (weakly and in mean square) at each $(s, t) \in T \times T$, and that $\hat{R}$ is strongly consistent (globally) in an appropriate norm.

1. Introduction. The likelihood ratio for the covariance of a Gaussian process of known mean has been studied a great deal, but primarily for the purpose of binary discrimination. For example, if an observed process $\{X_t, t \in T\}$ is of the form $X_t = S_t + N_t$, where S and N are unobservable, independent zero-mean Gaussian processes (signal and noise), with known covariance functions K_S and K_N, respectively, then we may test for the presence of the stochastic signal $\{S_t\}$ by deciding whether the covariance of $\{X_t\}$ is K_N or $K_S + K_N$. In this case we are confronted with two simple hypotheses, and whether we view this as a problem in testing or in discriminant analysis we are led to form the density dP_{S+N}/dP_N of the corresponding measures.

The likelihood ratio of two measures is a likelihood function evaluated at one point of the parameter space (which in this case is the set of all permissible covariances). In principal, we should be able to use the likelihood function to find the maximum likelihood estimate (MLE) of the covariance. For certain classes of covariances (all of them finitely parametrized), the MLE is known to exist almost surely. [Aside from the case $T = \{1, \ldots, n\}$ of multivariate analysis, this includes Anderson (1975), Section 5.4 of Anderson (1971), Azzalini (1981), Goodrich and Caines (1979), Hasza (1980), Hasza and Fuller (1979), Kashyap (1970) and Tugnait (1982) (see also Theorem 3.1).] Indeed, the only other cases in which likelihood methods have been applied turn out to involve families of mutually singular measures, for which perfect discrimination is possible [Bagchi (1975), Beder (1988), Grenander (1950), pages 221–222, Grenander (1981), pages 444–447, and Kelly, Reed and Root (1960), page 498].

Received February 1986; revised June 1987.

[1]Research supported in part by Air Force Office of Scientific Research Contract AFOSR 84-0329 and by a Research Incentive Program grant from the Graduate School of the University of Wisconsin, Milwaukee.

AMS 1980 *subject classifications*. Primary 62M09; secondary 60G15, 60G30.

Key words and phrases. Consistency, Gaussian dichotomy theorem, maximum likelihood estimation, reproducing kernel Hilbert space, sieve.

We will study this problem in the most general possible context. In particular, we will make no assumptions about the time-parameter set T. Thus, for example, our analysis will be applicable to random fields and to nonstationary processes. By using a convenient but general form of the likelihood, we will show that, for a sample of n replicates of the process, the MLE exists (almost surely) if a certain dimension is finite, and fails to exist (a.s.) if the dimension is infinite. This situation suggests the use of Grenander's theory of sieves [Geman and Hwang (1982) and Grenander (1981)], and, in fact, the finite-dimensional case will suggest an appropriate sieve. We will show that the resulting estimator is strongly consistent and give its exact distribution.

In the interest of brevity, certain proofs and background discussion will be omitted, but may be found in Beder (1987c). A more leisurely discussion of many of the ideas in this paper is given in a companion paper, Beder (1987a).

NOTATION AND DEFINITIONS. We will view a stochastic process as a family $\{X_t, t \in T\}$ of (real-valued) random variables defined on a measure space $(\Omega, \mathscr{A})$, where $\mathscr{A}$ is a σ-algebra. We will assume nothing about the set T.

Let V be the vector space of all finite linear combinations of the X_t. Under the probability measure Q on $(\Omega, \mathscr{A})$ this becomes a vector space V_Q of Q-equivalence classes of elements of V. We say that the process is *Gaussian* under Q if V_Q consists entirely of normal random variables. In this case, $V_Q \subset L^2(\Omega, \mathscr{A}, Q)$, and its completion $H_Q \subset L^2$ also consists of (possibly degenerate) normal random variables.

We denote the norm and inner product in H_Q by $\| \ \|_Q$ and $(\ , \ )_Q$, respectively. Expectation and covariance under Q are similarly denoted E_Q and Cov_Q.

One of our main tools is the reproducing kernel Hilbert space (RKHS) $\mathscr{H}(K, T)$, whose kernel $K(s, t)$ is a (real) positive symmetric function on $T \times T$ [Aronszajn (1950)]. In particular, if a second-order process has mean 0 and covariance R_Q under Q, then there is a natural isometry $\Lambda = \Lambda_Q$: $H_Q \to \mathscr{H}(R_Q, T)$, the *Loève map*, given by $Y \to g$, where $g(t) = (X_t, Y)_Q$ for all $t \in T$ [Loève (1948, 1977)].

Parzen (1959), Neveu (1968), Kallianpur (1970) and Kailath (1974) discuss the application of RKHS's to the study of Gaussian processes.

For ease of reading we will use the notation $f \otimes g$ for the function defined on $T \times T$ by $f \otimes g(s, t) = f(s)g(t)$. The notation $f \otimes g \otimes h$ will similarly represent the function on T^3 given by $f \otimes g \otimes h(s, t, u) = f(s)g(t)h(u)$; we will abbreviate $f \otimes f \otimes f \otimes f$ by $f^{4\otimes}$ (a function on T^4).

Finally, $\mathbb{Z}^+$ will denote the positive integers. It will be convenient to write

$$l_c^2(A) = \{\mathbf{a} \in l^2(A) : \inf\{a_\alpha\} > -1\}.$$

Usually A will be a finite set or $\mathbb{Z}^+$. This will be clear from context, and, in general, we will simply write l^2 and l_c^2.

2. The Gaussian dichotomy theorem (GDT) and its consequences. The Gaussian dichotomy theorem [see, e.g., Neveu (1968), pages 175–187] leads us to consider the model given by the largest set $\mathscr{P}$ of probability measures on $(\Omega, \mathscr{A})$

such that

(B_1) the process is Gaussian under every $Q \in \mathscr{P}$;
(B_2) the mean function of the process is identically zero under every $Q \in \mathscr{P}$;
(B_3) $\mathscr{P}$ is homogeneous; and
(B_4) the true probability measure belongs to $\mathscr{P}$.

Assumption (B_3) means that the measures in $\mathscr{P}$ are equivalent [Halmos and Savage (1949)]. We will arbitrarily single out a measure in $\mathscr{P}$ and denote it by P. In general, when a subscript is suppressed [e.g.,

$$H, (X, Y), \|X\|, \mathrm{Cov}(X, Y), E(X)],$$

we will understand it to be P. $\mathscr{H}$ will denote the RKHS $\mathscr{H}(R_p, T)$.

According to the GDT, for each $Q \in \mathscr{P}$ we have the following:

(i) There are a countable orthonormal set $\{g_k\}$ in $\mathscr{H}$ and a sequence $\mathbf{a} = \{a_k\} \in l_c^2$, both depending on Q, such that

$$R_Q(s, t) = R_p(s, t) + \sum a_k g_k(s) g_k(t), \quad \text{for all } s, t \in T. \tag{2.1}$$

Using the notation $f \otimes g$ given in the introduction, we may write the *parameter space* $\mathscr{C}$ (the set of covariances specified by the model) as

$$C = \left\{ R_Q = R_P + \sum a_k g_k \otimes g_k, \mathbf{a} \in l_c^2, \{g_k\} \text{ countable and orthonormal in } \mathscr{H} \right\}. \tag{2.2}$$

(ii) There are a countable orthonormal set $\{U_k\}$ in H and a sequence $\mathbf{b} = \{b_k\}$ satisfying $-\mathbf{b} \in l_c^2$, such that

$$\frac{dQ}{dp} = \exp 2^{-1} \sum \left(b_k U_k^2 + \ln(1 - b_k) \right). \tag{2.3}$$

Here

$$(1 + a_k)(1 - b_k) = 1 \quad \text{and} \quad g_k = \Lambda U_k, \quad \text{for all } k. \tag{2.4}$$

(iii) H and H_Q [and therefore $\mathscr{H}$ and $\mathscr{H}(R_Q, T)$] are isometric as Hilbert spaces and equal as sets. In particular, they have the same dimension, which may be uncountable; this dimension, as well as the set of observables H and the set of functions $\mathscr{H}$, depend only on $\mathscr{P}$, not on P and Q.

(iv) Let $Q^{n\otimes}$ denote the corresponding product measure on $(\Omega^n, \mathscr{A}^{n\otimes})$. Then we have

$$\frac{dQ^{n\otimes}}{dP^{n\otimes}} = \exp\left(\frac{n}{2} \sum_k \left(b_k S_k^2 - \ln(1 - b_k) \right) \right), \tag{2.5}$$

where S_k is the *second sample moment* of U_k,

$$S_k^2 = \frac{1}{n} \sum_{i=1}^{n} U_{ki}^2. \tag{2.6}$$

Equation (2.5) gives the density based on n independent replicates of the process. (Note that the sets $\{U_k\}$ and $\{S_k^2\}$ depend on Q.)

(v) The sequence $\{U_k/\sqrt{1+a_k}\}$ is i.i.d. $n(0,1)$ under Q, and the corresponding variables

$$Z_k = \frac{nS_k^2}{1+a_k} \tag{2.7}$$

are i.i.d. $\chi^2(n)$ under $Q^{n\otimes}$.

We will let $\mathscr{P}^{(n)} = \{Q^{n\otimes}, Q \in \mathscr{P}\}$.

REMARK 2.1. Although a different likelihood function would arise by a different choice of the measure P for the "denominator" of (2.3) or (2.5), it is easy to see that the method of maximum likelihood is not affected by that choice.

3. MLEs. Given a sample of n realizations of the process, we wish to maximize (2.5) by fixing $\omega \in \Omega^n$ and allowing both the set $\{S_k^2, k \in A\}$ and the sequence $-\mathbf{b} \in l_c^2(A)$ to vary. Here A is either $\mathbb{Z}^+$ or a finite set. Let us first consider the maximization when $\{S_k^2, k \in A\}$ and the corresponding set $\{g_k\}$ are fixed. It will, in fact, be useful to consider subsets $B \subset A$ and perform the restricted maximization over $l_c^2(B)$, the set of sequences that are zero off B. The proof of the following is similar to that of Theorem 3.1 in Beder (1987a), and will be omitted.

THEOREM 3.1. *Let $\{S_k^2, k \in A\}$ be fixed and let $l_c^2(B)$ be defined as before. If B is finite, then the likelihood* (2.5) *may be maximized over $l_c^2(B)$ almost surely $(\mathscr{P}^{(n)})$, the maximum occurring at $\hat{\mathbf{b}} = \{\hat{b}_k\}$ given by*

$$\begin{aligned} \hat{b}_k &= 1 - S_k^{-2}, \qquad k \in B, \\ &= 0, \qquad \text{otherwise}, \end{aligned} \tag{3.1a}$$

corresponding to $\hat{R}$ given by (2.1) *with $\hat{\mathbf{a}} = \{\hat{a}_k\}$ satisfying*

$$\begin{aligned} \hat{a}_k &= S_k^2 - 1, \qquad k \in B, \\ &= 0, \qquad \text{otherwise}. \end{aligned} \tag{3.1b}$$

If B is infinite, then the likelihood is unbounded over $l_c^2(B)$ a.s. $\mathscr{P}^{(n)}$.

Now, letting not only $\mathbf{b}$ (resp. $\mathbf{a}$) but also the set $\{S_k^2, k \in A\}$ (resp. $\{g_k, k \in A\}$) vary in (2.5) [resp. (2.1)], we immediately have

COROLLARY 3.1. *If* $\dim \mathscr{H} = \infty$, *then the likelihood for the covariance is unbounded almost surely.*

REMARK 3.1. As noted earlier, the criterion $\dim \mathscr{H} = \infty$ does not depend on the arbitrary choice of $P \in \mathscr{P}$. The almost sure unboundedness in the corollary must be handled with measure-theoretic care; see Beder (1987a), page 71.

4. The sieve estimator. The unboundedness of the likelihood function for the covariance leads us naturally to consider the method of sieves. We will use the following definition. Let $\mathscr{P} = \{P_\theta, \theta \in \Theta\}$ be a dominated family of probability measures (so that densities exist).

DEFINITION 4.1. A *sieve in* Θ is a collection $\{\mathscr{S}_m\}$ of subsets of Θ indexed by a parameter m such that

(a) $m' > m \Rightarrow \mathscr{S}_{m'} \supset \mathscr{S}_m$,
(b) $\cup \mathscr{S}_m$ is dense in Θ, and
(c) the likelihood can be maximized over each $\mathscr{S}_m$ (for some sample size n).

The restricted MLE $\hat{\theta} = \hat{\theta}_{mn}$ over each $\mathscr{S}_m$ for a sample of size n is called a *sieve estimator* of θ.

Our parameter space $\Theta = C$ (2.2) is a hugè set, involving not only the coefficients $\mathbf{a}$ but also the choice of orthonormal set $\{g_k\}$ in $\mathscr{H}$. A sieve in C is at present unavailable, but we will instead consider a sieve in a set $C_0 \subset C$ given by a *fixed, countably infinite* orthonormal set $\{g_k\}$ in $\mathscr{H}$,

$$C_0 = \left\{ R_Q \in C: R_Q = R_P + \sum_{k=1}^{\infty} a_k g_k \otimes g_k, \mathbf{a} \in l_c^2 \right\}. \tag{4.1}$$

If $\mathscr{H}$ is separable, we may make C_0 as large as possible by letting the set $\{g_k\}$ be complete in $\mathscr{H}$. Note that $R_P \in C_0$.

Define the sets

$$\mathscr{S}_m = \{\mathbf{a} \in l_c^2: a_k = 0 \text{ for } k > m\}, \qquad m = 1, 2, 3, \ldots . \tag{4.2}$$

It follows from Theorem 3.1 that $\{\mathscr{S}_m, m \in \mathbb{Z}^+\}$ is a sieve in l_c^2, where the resulting sieve estimator is

$$\hat{\mathbf{a}}_{mn} = (S_1^2 - 1, \ldots, S_m^2 - 1, 0, 0, \ldots); \tag{4.3}$$

here S_k^2 is defined by (2.6) with $U_k = \Lambda^{-1} g_k$. The sieve estimator in C_0 is now given by

$$\hat{R} = R_p + \sum_{k=1}^{m} \hat{a}_k g_k \otimes g_k. \tag{4.4}$$

We will develop the properties of $\hat{\mathbf{a}}$ and $\hat{R}$ under the assumption that

(C) the true covariance belongs to C_0.

The set C_0 is parametrized in one-to-one manner by l_c^2. Corresponding to C_0 is a set of measures $\mathscr{P}_0 \subset \mathscr{P}$, also parametrized by l_c^2. Thus (C) is the assumption that the true measure is in $\mathscr{P}_0$.

REMARK 4.1. If we wish to specify *two* covariances, say R_P and R_Q, which are to belong to the set of candidates C_0, then we must solve (2.1) for the sequences $\{g_k\}$ and $\mathbf{a}$; see Beder (1987b). If we only specify one, then $\{g_k\}$ can be constructed in innumerable ways, such as by the Gram–Schmidt process or (in certain cases) by diagonalizing an integral operator as in Cartier (1981), page 295.

Assumption (C) is discussed in Section 6.

We can describe the distribution of $\hat{R}(s, t)$ under assumption (C) by considering $\hat{R}$ as a stochastic process on $T \times T$. For simplicity, we will assume that dim $\mathscr{H}$ is countable [see (iii) of Section 2]; the inseparable case requires a minor elaboration. Extending $\{g_k, k \in \mathbb{Z}^+\}$ if necessary, we may assume that it is a complete orthonormal set in $\mathscr{H}$. Then a straightforward application of Section 1(v) along with Halmos (1967), problem 30, yields

THEOREM 4.1. *For each $m \in \mathbb{Z}^+$ and for any $\mathbf{a} \in l_c^2$, we have*

$$\hat{R} = (1/n) \sum_{k=1}^{m} Z_k(1 + a_k) g_k \otimes g_k + \sum_{k>m} g_k \otimes g_k, \tag{4.5}$$

where the variables Z_k are defined by (2.7). *In particular, under $Q \in \mathscr{P}_0$ given by* (2.1), *we have*

$$E_Q(\hat{R}) = R_P + \sum_{k=1}^{m} a_k g_k \otimes g_k \tag{4.6}$$

and

$$\operatorname{Cov}_Q(\hat{R}(s, t), \hat{R}(s', t')) = (2/n) \sum_{k=1}^{m} (1 + a_k)^2 g_k^{4\otimes}(s, t, s', t'). \tag{4.7}$$

REMARK 4.2. The random variables Z_k depend on m and n. Equations (4.5) and (4.6) are to hold at every $(s, t) \in T \times T$.

From this theorem we can easily deduce the following pointwise result. In the next section we will establish strong (and "global") consistency in an appropriate norm.

COROLLARY 4.1. *For $Q \in \mathscr{P}_0$ and for each $(s, t) \in T \times T$, $\hat{R}(s, t)$ is asymptotically unbiased as $m \to \infty$, and is weakly and mean-square consistent for $R_Q(s, t)$ if in addition we have $m = O(n)$.*

PROOF. Asymptotic unbiasedness follows immediately from (4.6).

For consistency, it suffices to show that $\operatorname{Var}_Q(\hat{R}(s, t)) \to 0$ as $m \to \infty$ if $m = O(n)$. But from (4.7) we have

$$\operatorname{Var}_Q \hat{R}(s, t) = (2/n) \sum_{k=1}^{m} (1 + a_k)^2 (g_k(s) g_k(t))^2. \tag{4.8}$$

Now $\Sigma g_k(s) g_k(t)$ converges for every $s, t \in T$ [in fact, it equals $R_P(s, t)$], so we have $(g_k(s) g_k(t))^2 \to 0$ as $k \to \infty$, whereas $\Sigma a_k^2 < \infty$ implies that $(1 + a_k)^2 \to 1$ as $k \to \infty$. Thus the summands in (4.8) go to 0 as $k \to \infty$, and an application of Lemma 4.1 shows that $\operatorname{Var}_Q \hat{R}(s, t) \to 0$. □

The lemma needed in the proof of Corollary 4.1 is an easy application of the Toeplitz lemma [Loève (1977), page 250], and will be useful later.

LEMMA 4.1. *Let* $x'_n = n^{-1}\sum_{k=1}^{m} x_k$ *for sequences* $\{x_n\}$ *and* $\{m\} = \{m_n\}$. *If* $x_n \to 0$ *and* $m = O(n)$ *as* $n \to \infty$, *then* $x'_n \to 0$. *If* $x_n \to x$ *and* $m/n \to \beta$ *as* $n \to \infty$, *then* $x'_n \to \beta x$.

5. Consistency in l^2. We continue to use the notation of the previous section, based on assumption (C) with a fixed orthonormal set $\{g_k, k \in \mathbb{Z}^+\}$ in $\mathscr{H}$. Parametrizing C_0 by l_c^2, we wish to show that $\|\hat{\mathbf{a}}_{mn} - \mathbf{a}\| \to 0$ a.s. $(P_{\mathbf{a}})$ for all $\mathbf{a} \in l_c^2$, as long as m increases at an appropriate rate with n.

Let us fix $\mathbf{a} \in l_c^2$; then

$$\|\hat{\mathbf{a}}_{mn} - \mathbf{a}\|^2 = \sum_{k=1}^{m} (\hat{a}_{mnk} - a_k)^2 + \sum_{k>m} a_k^2$$
$$= W_{mn} + \sum_{k>m} a_k^2, \quad \text{say.}$$

Clearly, we must have $m \to \infty$ for the latter term to vanish. The stochastic term W_{mn} is positive, and it is easy to see that

$$W_{mn} = (1/n^2) \sum_{k=1}^{m} (1 + a_k)^2 (Z_k - n)^2, \tag{5.1}$$

where the variables Z_k are defined by (2.7). We thus have the following weak result, the latter part of which rests on Chebyshev's inequality, Lemma 4.1 and the fact that $a_k \to 0$.

LEMMA 5.1. *Under* $P_{\mathbf{a}}$, *the distribution of* W_{mn} *is given by* (5.1), *where the variables* Z_k *are i.i.d.* $\chi^2(n)$. *In particular, if* $m/n \to \beta < \infty$ *as* $n \to \infty$, *then* $W_{mn} \to 2\beta$ *in* $P_{\mathbf{a}}$*-probability.*

This result sets "best possible" limits on strong consistency results. In particular, if we choose m so that $m/n \to \beta$, then it will certainly be necessary to have $\beta = 0$, that is, $m = o(n)$. Note that the pointwise consistency result of Corollary 4.1 required only that $m = O(n)$.

Now strong consistency is equivalent to the condition that for every $\varepsilon > 0$ we have

$$P_{\mathbf{a}}(W_{mn} > \varepsilon \text{ i.o.}) = 0, \tag{5.2}$$

since W_{mn} is nonnegative. From the Borel–Cantelli lemma, then, it suffices to pick $m = m_n$ so that for every $\varepsilon > 0$ we have

$$\sum_{n=1}^{\infty} P_{\mathbf{a}}(W_{mn} > \varepsilon) < \infty. \tag{5.3}$$

First let us get a simple bound on each term in (5.3).

LEMMA 5.2. *For each* $\varepsilon > 0$ *and for* $r = 1, 2, 3, \ldots$, *we have* $P_{\mathbf{a}}(W_{mn} > \varepsilon) < c_{rmn}(\varepsilon)$, *where*

$$c_{rmn}(\varepsilon) = \left[(4/n\varepsilon) \sum_{k=1}^{m} (1 + a_k)^2\right]^r (2r)!(n/(n-1))^m. \tag{5.4}$$

The extra "parameter" r will give us the necessary leverage to force $\{c_{rmn}(\varepsilon)\}$ to be summable and thus (5.3) to hold. We will, in fact, let r as well as m depend on n. The proof of Lemma 5.2 rests on the following fact, which is derived from problem 16.3 of Kendall and Stuart (1969).

LEMMA 5.3. *Let Z be $\chi^2(n)$, $n \geq 2$, and let s be a nonnegative integer. Then*

$$E\left[(Z-n)^{2s}\right] < (4n)^s(2s)!\, n/(n-1). \tag{5.5}$$

PROOF OF LEMMA 5.2. By Markov's inequality we have, for every r,

$$\begin{aligned} P_{\mathbf{a}}(W_{mn} > \varepsilon) &= P_{\mathbf{a}}\left(\left(n^2W_{mn}\right)^r > \left(n^2\varepsilon\right)^r\right) \\ &\leq \frac{E\left[\left(n^2W_{mn}\right)^r\right]}{\left(n^2\varepsilon\right)^r} \end{aligned} \tag{5.6}$$

(expectation taken under $P_{\mathbf{a}}$). To get a bound for the numerator, let us rewrite (5.1) for simplicity as

$$n^2W_{mn} = \sum_{k=1}^{m} d_kY_k^2,$$

where $d_k = (1 + a_k)^2$ and $Y_k = Z_k - n$. Now the Y_k are independent, and so for every $r \in \mathbb{Z}^+$ we have

$$E\left[\left(n^2W_{mn}\right)^r\right] = \Sigma\binom{r}{r_1 \cdots r_m} d_1^{r_1} \cdots d_m^{r_m} E\left(Y_1^{2r_1}\right) \cdots E\left(Y_m^{2r_m}\right), \tag{5.7}$$

where the sum is over all m-tuples $(r_1, \ldots, r_m)$ which sum to r and satisfy $r_k \geq 0$ for each k. But $Y_k = Z_k - n$ is a centered χ^2 random variable with n d.f. under $P_{\mathbf{a}}$, so from Lemma 5.3 we have

$$\begin{aligned} E\left(Y_1^{2r_1}\right) \cdots E\left(Y_m^{2r_m}\right) &< (4n)^r\left(\frac{n}{n-1}\right)^m \prod_{k=1}^{m} (2r_k)! \\ &\leq (4n)^r\left(\frac{n}{n-1}\right)^m (2r)!, \end{aligned} \tag{5.8}$$

the second inequality due to the fact that the ratio $(2r)!/\prod(2r_k)!$ is a multinomial coefficient and so is at least 1. Substituting (5.8) into (5.7), we get

$$\begin{aligned} E\left(n^2W_{mn}\right)^r &< (4n)^r\left(\frac{n}{n-1}\right)^m (2r)! \sum\binom{r}{r_1 \cdots r_m} d_1^{r_1} \cdots d_m^{r_m} \\ &= (4n)^r\left(\frac{n}{n-1}\right)^m (2r)!\left(\sum d_k\right)^r \quad \text{(summing over } k = 1, \ldots, m) \\ &= \left(4n\sum d_k\right)^r\left(\frac{n}{n-1}\right)^m (2r)!, \end{aligned}$$

and substituting *this* into (5.6) finally yields (5.4). □

Our goal now is to choose $r = r_n$ and $m = m_n$ so that the bounds $c_{rmn}(\varepsilon)$ form a summable sequence for each $\varepsilon > 0$. While our interest is in choosing $m = o(n)$, that order of growth will not yield summability by the present method of proof, and we must instead require that $m = o(n^\sigma)$ for some $\sigma \in (0,1)$. Actually, it will cost us nothing to consider m such that $m/n^\sigma \to \beta \geq 0$.

LEMMA 5.4. *Suppose* $\{r_n\}$ *and* $\{m_n\}$ *are chosen so that for some* $\sigma, \tau > 0$ *such that* $\sigma + 2\tau = 1$, *we have*

(i) $m/n^\sigma \to \beta \geq 0$ *and*
(ii) $(r/n^\tau)^r n^{1/4}$ *is summable and* $r \geq n$ *for* n *large.*

Then the bounds $c_{rmn}(\varepsilon)$ *given by* (5.4) *form a summable sequence for all* $\varepsilon \geq 16\beta/e^2$.

PROOF. We continue using the notation $d_k = (1 + a_k)^2$; note that since a_k is square summable we have $d_k \to 1$ as $k \to \infty$.

Fix ε, and let $c_n = c_{rmn}(\varepsilon)$. First, condition (i) implies that $(n/(n-1))^m \to 1$ as $n \to \infty$, so that

$$c_n \sim \left[\frac{4\Sigma d_k}{n\varepsilon}\right]^r (2r)!,$$

where $c_n \sim b_n$ means that $c_n/b_n \to 1$, and summation is over $k = 1, \ldots, m$. By Stirling's formula, $(2r)! \sim 2\sqrt{\pi}\,(4r^2/e^2)^r r^{1/2}$, so that

$$c_n/2\sqrt{\pi} \sim \left(\frac{16r^2\Sigma d_k}{ne^2\varepsilon}\right)^r r^{1/2},$$

which we write as

$$\left(\frac{16}{e^2\varepsilon}\frac{1}{n^\sigma}\Sigma d_k\right)^r \left(\frac{r}{n^\tau}\right)^{2r} r^{1/2}. \tag{5.9}$$

From Lemma 4.1 and condition (i) we see that $(1/n^\sigma)\Sigma d_k \to \beta$ as $n \to \infty$, so that the first factor in (5.9) is bounded by $1^r = 1$ for large n as long as $16\beta/e^2 \leq \varepsilon$. Thus $c_n/2\sqrt{\pi}$ is summable by comparison with $(r/n^\tau)^{2r} r^{1/2}$, which is summable by comparison with the sequence in (ii). □

There are many sequences satisfying condition (ii) of Lemma 5.4; see Examples A.1 and A.2 in the Appendix. Thus we have established the following.

THEOREM 5.1. *Let* $m \to \infty$ *and* $m/n^\sigma \to \beta < \infty$ *for some* $\sigma \in (0,1)$. *Then for any* $\varepsilon > 16\beta/e^2$ *we have* $P_{\mathbf{a}}(W_{mn} < \varepsilon \text{ i.o.}) = 1$.

From this we immediately get our main result.

COROLLARY 5.1. *If* $m \to \infty$ *and* $m/n^\sigma \to 0$ *for some* $\sigma \in (0,1)$, *then* $\|\hat{\mathbf{a}}_{mn} - \mathbf{a}\| \to 0$ *a.s.* $P_{\mathbf{a}}$.

In other words, $\hat{\mathbf{a}}$ is strongly l^2-consistent for $\mathbf{a}$, for every $\mathbf{a} \in l_c^2$, and so for every covariance $R \in C_0$. The rate $m = o(n^\sigma)$ of Corollary 5.1 is almost best possible in the sense we have discussed. On the other other hand, if $m/n^\sigma \to \beta > 0$, so that we do not have even weak convergence, we can still gain some insight into what is happening.

COROLLARY 5.2. *If $m/n^\sigma \to \beta > 0$ for some $\sigma \in (0,1)$, then W_{mn} is almost surely bounded by ε when n is large, for any $\varepsilon \geq 16\beta/e^2$, and the square error $\|\hat{\mathbf{a}} - \mathbf{a}\|^2$ is almost surely bounded as $n \to \infty$.*

REMARK 5.1. An alternative method of proof [see Antoniadis and Beder (1988)] does show that $\sigma = 1$ is attainable. This is certainly not a matter of direct practical importance, since the growth rate of m is merely a guide to the proper choice of sieve size m_n for a given sample size n, and in any finite experiment it would be impossible to distinguish $m = o(n)$ from $m = o(n^{0.99})$, say. However, it *is* of theoretical interest that when the parameter space is (contained in) a Hilbert space and m is the dimension of a subspace, the method of sieves seems to produce $m = o(n)$ as an optimal rate [see McKeague (1986), page 580, and Beder (1987a)].

6. Interpreting the metric. Grenander (1981), pages 444–446, suggests studying the covariance estimation problem by considering operator norms, but under very strong assumptions about sample paths and about the family of covariances (in fact, in his case a sieve is not even necessary, as he points out). The present approach allows us to avoid his assumptions. At the same time, the distance $\|\hat{\mathbf{a}} - \mathbf{a}\|$ used in Section 5 has a natural interpretation as a distance between covariances. We state the main result here without proof.

THEOREM 6.1. *Let R_1 and R_2 be two covariances in C_0, corresponding to sequences $\mathbf{a}_1$ and $\mathbf{a}_2 \in l_c^2$. Then $\|\mathbf{a}_1 - \mathbf{a}_2\| = \|R_1 - R_2\|$, where the first norm is that of l^2 and the second is that of $\mathscr{H}^{2\odot}$.*

Here $\mathscr{H}^{2\odot}$ is the second symmetric tensor power, or second Wiener chaos, of $\mathscr{H} = \mathscr{H}(R_P, T)$. [The general theory of tensor powers is given in Neveu (1968); see also Guichardet (1972) and Kallianpur (1980).]

Theorem 6.1 applies in particular to $R_1 = R_Q$ and $R_2 = \hat{R}$, say, where $\hat{R}$ is a sieve estimator in C_0 and $R_Q \in C_0$. Since the norm in $\mathscr{H}^{2\odot}$ is coordinate-free, it might be useful in eliminating assumption (C), which depends upon our fixing a coordinate system $\{g_k\}$ in $\mathscr{H}$. But this remains to be seen.

Using an isometry given in Beder (1987b), the norm given previously can sometimes be interpreted as a norm in $L^2(T^2)$, bringing it closer to Grenander's original idea.

7. Conclusion. We have constructed an estimator of the covariance of a general Gaussian process based on n realizations of the process. This estimator is straightforward to compute, is analytically tractable and behaves well

asymptotically. The price we have paid for our generality is the use of replicates and the assumption (C) that the true covariance belongs to a subset C_0 of the original parameter space C.

Unlike stationarity, which by its very nature allows us to reuse a single trajectory to estimate the covariance at different lags, nonstationarity of a process would seem to require us to observe several trajectories (replicates) of the process for covariance estimation. The method proposed here certainly rests on the availability of such data. Although not an assumption in the same sense as our assumptions (B1)–(B4) and (C), this *is* a design constraint. A discussion of its use in both the theoretical and the applied literature is given in Beder (1987c).

Finally, our consistency results leave open the questions of the *rate* of convergence of our estimator, and of the optimal choice and ordering of the orthonormal set $\{g_k\}$ in $\mathscr{H}$. These are typical problems in sieve theory, and we can expect the availability of an exact distribution theory to aid us in investigating them.

APPENDIX

In this section we give some examples of sequences $\{r_n\}$ of the type needed in Section 5.

The summability of a sequence of form $(r/n^\tau)^r$, where $r = r_n \to \infty$, and more generally of $(r/n^\tau)^r n^\gamma$, is somewhat delicate. If r grows too fast, then the terms r/n^τ may grow so fast that $(r/n^\tau)^r$ fails to go to 0, whereas if r grows too slowly, r/n^τ may go to 0 at a reasonable rate but $(r/n^\tau)^r$ will fail to decrease fast enough [note that it is *not* bounded by the summable sequence $(r/n^\tau)^n$].

Let us say that the sequence r is *in class* (τ, α), $\tau > 0$, $\alpha \geq 0$, if as $n \to \infty$ we have

(i) $r/n^\tau \to 0$, and
(ii) $n^\alpha \delta^r \to 0$ for some $\delta > 0$.

Of course, if (ii) holds for some δ_0, then it holds for all $\delta < \delta_0$.

LEMMA A.1. *If the sequence r is in class (τ, α), then $r \to \infty$ and $(r/n^\tau)^r = o(n^{-\alpha})$ as $n \to \infty$. In particular, if $\alpha - \gamma > 1$ and if r is in class (τ, α), then $(r/n^\tau)^r n^\gamma$ is summable.*

PROOF. The fact that $r \to \infty$ is immediate from (ii), taking $\delta < 1$.

From (i) we see that for *all* $\delta > 0$, we have $r/n^\tau < \delta$ for n large, so that $n^\alpha (r/n^\tau)^r < n^\alpha \delta^r$ for n large. Thus condition (ii) implies that $n^\alpha (r/n^\tau)^r \to 0$ as $n \to \infty$, as desired.

Finally, summability is seen to hold by comparison with the sequence $n^{-(\alpha-\gamma)}$. □

EXAMPLE A.1. If $r = n^\beta, 0 < \beta < 1$, then r is in class (τ, α) for $\tau > \beta$ and for every α. Condition (i) is easy to see, whereas (ii) holds if we take $\delta < 1$. Thus

the sequence

$$\left(\frac{n^\beta}{n^\tau}\right)^{n^\beta} = n^{(\beta-\tau)n^\beta}$$

is summable, and so is

$$\left(\frac{n^\beta}{n^\tau}\right)^{n^\beta} n^\gamma = n^{(\beta-\tau)n^\beta+\gamma},$$

for any $\gamma > 0$, as long as $\tau > \beta$.

EXAMPLE A.2. If $r = \ln n$, then r is in class (τ, α) for every τ and α. Here again condition (i) is obvious, whereas (ii) holds for any given $\alpha > 0$ if we take $\delta < e^{-\alpha}$. Thus such sequences as

$$\left(\frac{\ln n}{n^\tau}\right)^{\ln n} n^\gamma$$

are summable for any $\gamma > 0$.

Acknowledgments. I would like to thank the referees for their helpful and insightful comments. I was also greatly aided by discussions with a number of people, including Anestis Antoniadis, Arun Garg, Stuart Geman, Jugal Ghorai and Ian McKeague. Finally, I would like to add a special note of gratitude to Robert Shumway, who not only contributed valuable suggestions, but also directed my dissertation, out of which grew Sections 2 and 3 of this paper.

REFERENCES

ANDERSON, T. W. (1971). *The Statistical Analysis of Time Series*. Wiley, New York.

ANDERSON, T. W. (1975). Maximum likelihood estimation of parameters of autoregressive processes with moving average residuals and other covariance matrices with linear structure. *Ann. Statist.* **3** 1283–1304.

ANTONIADIS, A. and BEDER, J. H. (1988). Joint estimation of the mean and the covariance of a Banach-valued Gaussian vector. *Statistics*. To appear.

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404.

AZZALINI, A. (1981). Replicated observations of low order autoregressive time series. *J. Time Ser. Anal.* **2** 63–70.

BAGCHI, A. (1975). Continuous time systems identification with unknown noise parameters. *Automatica* **11** 533–536.

BEDER, J. H. (1987a). A sieve estimator for the mean of a Gaussian process. *Ann. Statist.* **15** 59–78.

BEDER, J. H. (1987b). Simultaneous diagonalization of two covariance kernels. Unpublished.

BEDER, J. H. (1987c). A sieve estimator for the covariance of a Gaussian process: Theory and background. Unpublished.

BEDER, J. H. (1988). Estimating a covariance function having an unknown scale parameter. *Comm. Statist. A—Theory Methods* **17** 323–340.

CARTIER, P. (1981). Une étude des covariances mesurables. In *Mathematical Analysis and Applications, Part A* (L. Nachbin, ed.) 267–316. Academic, New York.

GEMAN, S. and HWANG, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414.

GOODRICH, R. L. and CAINES, P. E. (1979). Linear system identification from nonstationary cross-sectional data. *IEEE Trans. Automat. Control* **AC-24** 403–411.

GRENANDER, U. (1950). Stochastic processes and statistical inference. *Ark. Mat.* **1** 195–277.

GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York.

GUICHARDET, A. (1972). *Symmetric Hilbert Spaces and Related Topics. Lecture Notes in Math.* **261** Springer, New York.

HALMOS, P. R. (1967). *A Hilbert Space Problem Book*. Van Nostrand-Reinhold, Princeton, N.J.

HALMOS, P. R. and SAVAGE, L. J. (1949). Application of the Radon–Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Statist.* **20** 225–241.

HASZA, D. P. (1980). A note on maximum likelihood estimation for the first order autoregressive process. *Comm. Statist. A—Theory Methods* **9** 1411–1415.

HASZA, D. P. and FULLER, W. A. (1979). Estimation for autoregressive processes with unit roots. *Ann. Statist.* **7** 1106–1120.

KAILATH, T. (1974). A view of three decades of linear filtering theory. *IEEE Trans. Inform. Theory* **IT-20** 146–181.

KALLIANPUR, G. (1970). The role of reproducing kernel Hilbert spaces in the study of Gaussian processes. In *Advances in Probability* (P. Ney, ed.) 49–83. Dekker, New York.

KALLIANPUR, G. (1980). *Stochastic Filtering Theory*. Springer, New York.

KASHYAP, R. L. (1970). Maximum likelihood identification of stochastic linear systems. *IEEE Trans. Automat. Control* **AC-15** 25–34.

KELLY, E. J., REED, I. S. and ROOT, W. L. (1960). The detection of radar echoes in noise. II. *J. SIAM* **8** 481–507.

KENDALL, M. G. and STUART, A. (1969). *The Advanced Theory of Statistics* **1**. Hafner, New York.

LOÈVE, M. (1948). Fonctions aléatoires du second ordre. Supplement to P. Lévy, *Processus Stochastiques et Mouvement Brownien*. Gauthier-Villars, Paris.

LOÈVE, M. (1977). *Probability Theory* **1**, 4th ed. Springer, New York.

MCKEAGUE, I. (1986). Estimation for a semimartingale regrression model using the method of sieves. *Ann. Statist.* **14** 579–589.

NEVEU, J. (1968). *Processus Aléatoires Gaussiens*. Publications du Seminaire de Mathématiques Supérieures. Les Presses de l'Université de Montréal, Montréal.

PARZEN, E. (1959). Statistical inference on time series by Hilbert methods. I. Technical Report 23, Dept. Statistics, Stanford Univ. Reprinted in *Time Series Analysis Papers*, paper 13. Holden-Day, San Francisco (1967).

TUGNAIT, J. K. (1982). Global identification of continuous-time systems with unknown noise covariance. *IEEE Trans. Inform. Theory* **IT-28** 531–536.

DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF WISCONSIN
P.O. BOX 413
MILWAUKEE, WISCONSIN 53201