The Annals of Statistics
1987, Vol. 15, No. 1, 59–78

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

BY JAY H. BEDER

University of Wisconsin-Milwaukee

A new sieve estimator for the mean function $m(t)$ of a general Gaussian process of known covariance is presented. The estimator $\hat{m}(t)$ is given explicitly from the data and has a simple distribution. It is shown that $\hat{m}(t)$ is asymptotically unbiased and consistent (weakly and in mean square) at each t, and that $\hat{m}$ is strongly consistent for m in an appropriate norm. No assumptions are made about the "time" parameter or the covariance.

1. Introduction.

It was observed at least as early as 1968 that the method of maximum likelihood fails in general to give an estimate of the mean of a Gaussian process "in an infinite-dimensional case" [Rozanov (1971), page 128; special instances were known earlier]. Roughly speaking, the parameter space is "too large" and consequently the likelihood function is unbounded.

Grenander (1981) introduced the method of sieves as a means of rescuing maximum likelihood in just such cases. In fact, he provided an example of sieve estimation of the mean of a Wiener process of known covariance [(1981), pages 422–424; see the example below], which Geman and Hwang [(1982), pages 409–411] used to illustrate a general consistency theorem: the sieve estimator converges (in a certain norm) almost surely to the true mean. The construction of this estimator does not depend on the particular covariance used, and will be generalized to arbitrary covariances in the following. Due to the nature of the sieve, however, their estimator depends on a random variable λ which is defined only implicitly from the data. Aside from the computational problem this creates, it also means that small-sample properties of the estimator will be difficult to discover.

The sieve estimator that we will consider is given explicitly in terms of independent normal random variables, and so its distribution can be written down precisely. Its strong-consistency proof involves elementary Chebyshev bounds and a straightforward application of the Borel–Cantelli lemma.

An example. To illustrate our results, we will frequently turn to the example of a standard Wiener process, which is essentially that used by Grenander and by Geman and Hwang.

Received December 1985; revised April 1986.

[1] Research was 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.

The standard Wiener process on $T = [0, b]$ is a Gaussian process $\{X_t, t \in T\}$ with covariance $R(s, t) = \min(s, t)$. We will assume that the mean function m is unknown and must be estimated based on n realizations of the process (as we shall say, a sample of size n). Thus we consider the set $\mathscr{P}$ of probability measures under which the process is Gaussian with covariance R and with some mean function $m(t)$. $\mathscr{P}$ is of course indexed, or parametrized, by the set of mean functions, among which we admit the function $m(t) \equiv 0$. As we will note in the next section, each mean function may be written in the form

$$m(t) = \int_0^t \gamma(u)\,du,$$

for a unique $\gamma \in L^2(T)$, and each γ gives rise to a mean function m. Thus $L^2(T)$ is a parameter space for the model, and is the one used by Grenander.

Estimating m is equivalent to estimating γ. But γ in turn may be written in an orthogonal expansion

$$\gamma(t) = \sum a_k \gamma_k(t), \tag{1.1}$$

where $\{\gamma_k\}$ is a complete orthonormal (CON) basis of $L^2(T)$. For a fixed CON basis, the Fourier coefficients a_k in (1.1) are unique to γ and the sequence $\mathbf{a} = \{a_k\}$ is square summable, so that we in fact reparametrize our model, this time by the set ℓ^2. Now we must estimate $\mathbf{a}$. But the likelihood $L(\mathbf{a})$ is seen to be almost surely unbounded over ℓ^2 (see Section 3 below).

Grenander's sieve estimator in ℓ^2 is the restricted maximum likelihood estimator (MLE) $\hat{\mathbf{a}}$ which, for a sample of size n, maximizes L over the "ellipsoid"

$$\mathscr{S}_d = \left\{\mathbf{a} \in \ell^2 \colon \sum k^2 a_k^2 \le d\right\}, \tag{1.2}$$

where d is a positive real. Note that $\hat{\mathbf{a}}$ depends on both d and n. (The *sieve* is the collection of these sets $\mathscr{S}_d$.) The corresponding estimator $\hat{\gamma}$ in L^2 is then given by

$$\hat{\gamma}(t) = \sum \hat{a}_k \gamma_k(t).$$

Geman and Hwang showed how to choose d as a function of n so that $\|\hat{\mathbf{a}} - \mathbf{a}\| \to 0$ (ℓ^2-norm) with $P_{\mathbf{a}}$-probability 1 as $n \to \infty$, and thus so that $\|\hat{\gamma} - \gamma\| \to 0$ [$L^2(T)$-norm] almost surely as well. The difficulty, as already mentioned, is that $\hat{\mathbf{a}}$ cannot be gotten in explicit form.

We will replace Grenander's ellipsoids by the sets

$$\mathscr{S}_d = \left\{\mathbf{a} \in \ell^2 \colon a_k = 0 \text{ for } k > d\right\}, \tag{1.3}$$

where d is now a positive integer. These sets are of course subspaces of ℓ^2 of dimension d. The resulting estimator $\hat{\mathbf{a}}$ is far more tractable than the previous one, and our consistency proof will be less delicate than that used by Geman and Hwang.

It turns out that Grenander [(1981), page 424] also proposed the sieve given by (1.3), but did not develop it. For purposes of distinction, we will refer to the sieve defined by (1.2) as the GGH (Grenander–Geman–Hwang) sieve.

Nguyen and Pham (1982) use a sieve similar to that in (1.3) to estimate the parameter $\theta(t)$ in the diffusion model

$$dX(t) = \theta(t)X(t)\,dt + dW(t), \qquad t \geq 0, \tag{1.4}$$

where $W(t)$ is a zero-mean Wiener process, getting strong consistency in ℓ^2-norm when $d = o(n^{1/2})$. They point out that their methods can be adapted to estimating θ in the model

$$dX(t) = \theta(t)\,dt + dW(t),$$

which, when written

$$X(t) = m(t) + W(t), \qquad m(t) = \int_0^t \theta(u)\,du, \tag{1.5}$$

is precisely our example with mean function m. Carrying out their construction for (1.5) leads to the new sieve (1.3).

McKeague (1986) has used the method of sieves for more general models than (1.4), getting the improved (relaxed) growth rate $d = o(n)$ for strong consistency. We will generalize in a different direction, replacing $W(t)$ in (1.5) by a zero-mean Gaussian process on an arbitrary "time" set T, with arbitrary fixed covariance function R. To do this we will bypass the $L^2(T)$ parametrization, which can no longer be assumed to exist, and recognize that the set of candidate mean functions is always itself a Hilbert space $\mathscr{H} = \mathscr{H}(R, T)$, the reproducing kernel Hilbert space (RKHS) with kernel R. We will see that the condition $d = o(n)$ is best possible in this case.

With our RKHS formulation, we will also have shown that the consistency result of Geman and Hwang for the GGH sieve applies to an arbitrary Gaussian process of known covariance, with convergence equivalently in $\mathscr{H}$ or in ℓ^2. Antoniadis (1985) has generalized their work in another direction. A recent survey of these and other results is given by McKeague (1985). Finally, it may be noted that a sieve similar to (1.3) has been applied to estimating the covariance of a Gaussian process of known mean, in which the condition $d = o(n)$ again emerges as best possible [Beder (1986)].

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 $(\Omega, \mathscr{A})$. 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 the completion H_Q 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, and the *mean function* of the process under Q is given by

$$m_Q(t) = E_Q(X_t).$$

We denote by $\sigma(H_Q)$ the σ-algebra generated by the process and the sets of Q-measure zero.

The real numbers are denoted by $\mathbb{R}$, and the positive integers by $\mathbb{Z}^+$.

2. The Gaussian dichotomy theorem and its consequences. Our goals in this section are to describe the model $\mathscr{P}$ which we will be assuming, to specify its parametrizations, to write down its likelihood function and to derive some elementary distribution theory. Underlying all of this is the Gaussian dichotomy theorem, which we will now state.

The dichotomy of this theorem is the assertion that, under certain conditions, two measures P and Q must be singular (= orthogonal; denoted $P \perp Q$) or equivalent (= mutually absolutely continuous; denoted $P \sim Q$). A proof of the theorem, with no conditions on the index set T, is given by Neveu [(1968), pages 171–174]. We may restate the theorem in the following convenient form.

THEOREM 2.1. *Let $(X_t, t \in T)$ be a real-valued Gaussian process on $(\Omega, \mathscr{A})$ with covariance R under both P and Q, but with zero mean under P. Let $H = H_P$, and assume $\mathscr{A} = \sigma(H)$. Then the following are equivalent*:

$$\text{(2.1)} \quad \begin{array}{l} \text{(i) } P \not\perp Q. \\ \text{(ii) } E_Q(\cdot) = (\cdot, Y) \text{ for a unique } Y \in H. \end{array}$$

$$\text{(2.2)} \quad \begin{array}{l} \text{(iii) } dQ/dP = e^Y/E(e^Y) \text{ for a unique } Y \in H. \\ \text{(iv) } P \sim Q. \end{array}$$

If these conditions hold, then the random variable Y in (iii) *is the same as that in* (ii). *[$(\ ,\ ) = (\ ,\ )_P$ and $E = E_P$.]*

Conversely, every $Y \in H$ gives rise via (ii) *or* (iii) *to a probability measure Q on $(\Omega, \mathscr{A})$ which makes the process Gaussian with covariance R (and which is equivalent to P).*

CONSEQUENCE 1: THE MODEL. Let us consider the collection of all probability measures on $(\Omega, \mathscr{A})$ with respect to which the process is Gaussian with a fixed covariance R. These measures are distinguished by the mean functions with which they endow the process. Theorem 2.1 expresses the idea that this collection is a disjoint union of subfamilies, such that each subfamily $\mathscr{P}$ consists of equivalent measures and such that the measures in $\mathscr{P}$ are orthogonal to those outside of $\mathscr{P}$. [A family of equivalent measures is said to be *homogeneous*, a term due to Halmos and Savage (1949).] In principle, a single observation will enable us to decide precisely which subfamily contains the true probability measure which is generating the data.

Now the set of all candidate mean functions is similarly partitioned into subfamilies, and so in principle a single observation of the process will tell us which subfamily contains the true mean function. It has become a standard assumption that this subfamily has already been identified and so is known to us. The remaining problem is to infer which mean function in this family is the true one. A further reduction in the problem is to select a mean function from the given family and to subtract it from the process (this does not alter $\mathscr{A}$), so that the set of mean functions may be assumed to include the zero function. We will reserve the notation P for the corresponding measure. Theorem 2.1 thus leads us to consider the model given by the largest set $\mathscr{P}$ of probability measures on $(\Omega, \mathscr{A})$ such that

(A_1) the process is Gaussian under every $Q \in \mathscr{P}$;
(A_2) the covariance of the process is the same, say R, under every $Q \in \mathscr{P}$;
(A_3) $\mathscr{P}$ is homogeneous and $\mathscr{A} = \sigma(H_Q)$ for any $Q \in \mathscr{P}$; and
(A_4) there is a measure $P \in \mathscr{P}$ under which the process has mean zero.

Assumption A_2 says that for all s and $t \in T$ we have

$$\mathrm{Cov}_Q(X_s, X_t) = R(s,t), \quad \text{for all } Q \in \mathscr{P}. \tag{2.3}$$

For convenience, the subscript P will often be suppressed: $E_P = E$, $H_P = H$, etc. Distribution theory under P may be summarized rather conveniently:

Lemma 2.1. *Let X and Y be elements of H. Then (under P) their distribution is normal with mean zero. We have*

$$\mathrm{Cov}(X,Y) = (X,Y),$$

and, in particular,

$$\mathrm{Var}(X) = \|X\|^2.$$

Consequence 2: the parametrizations. Let

$$\mathscr{H} = \{m_Q, Q \in \mathscr{P}\}$$

be the set of mean functions defined by the model. From (2.1) we see that every mean function is of the form

$$m(t) = (X_t, Y), \tag{2.4}$$

for a unique $Y \in H$, and that every $Y \in H$ gives rise to a mean function in this way. From this it follows that $\mathscr{H}$ is a vector space under pointwise addition. The correspondence Λ given by

$$\Lambda(Y) = m \tag{2.5}$$

is the *Loève map* [Loève (1948)].

Corollary 2.1. *Λ maps H onto $\mathscr{H}$, and is a vector space isomorphism. Moreover, if $\mathscr{H}$ inherits an inner product $\langle\ ,\ \rangle$ from H via Λ, viz.,*

$$\langle g, h\rangle = (Y,Z), \quad \textit{where } g = \Lambda(Y) \textit{ and } h = \Lambda(Z), \tag{2.6}$$

then $\mathscr{H}$ becomes the reproducing kernel Hilbert space $\mathscr{H}(R,T)$ with kernel R, and Λ is an isometry.

This result is well known; our formulation follows Neveu [(1968), pages 34–36]. For us the important points are:

(1) the set of mean functions is a Hilbert space; and
(2) the Loève map gives a direct way to compute a mean function from a density and vice-versa. More precisely, if a density is of form

$$dQ/dP = e^{Y}/E(e^{Y})$$

as in (2.2), then the function $m = \Lambda(Y)$ is the mean of the process under the measure Q; conversely, if a mean function m is given corresponding to a measure Q, then the density dQ/dP is of the above form with $Y = \Lambda^{-1}(m)$.

Remark 2.1. Corollary 2.1 also provides a criterion for the equivalence of the measures P and Q; namely, $P \sim Q$ iff $m_Q \in \mathscr{H}$. This was first recognized by Parzen [(1959), Theorem 9A] and by Kallianpur and Oodaira (1963), although under some restrictions.

Remark 2.2. We are assuming some familiarity with the idea of a reproducing kernel Hilbert space (RKHS). The basic paper in this area is Aronszajn (1950). Sources dealing with the role of RKHS's in the study of second-order processes include Parzen (1959), Neveu (1968) and Kallianpur (1970). The most important facts are these. First, $\mathscr{H} = \mathscr{H}(R, T)$ satisfies two properties:

(a) for each $t \in T$ the function R_t defined by $R_t(s) = R(s, t)$ belongs to $\mathscr{H}$;
(b) (reproducing property) if $h \in \mathscr{H}$ and $t \in T$, then $\langle h, R_t\rangle = h(t)$.

We see immediately that the functions R_t span $\mathscr{H}$ in the sense that if $h \perp R_t$ for all $t \in T$, then $h = 0$. Second, from (2.4) we see that $\Lambda(X_t) = R_t$. Evaluating the Loève map or its inverse is crucial, since as we noted above it connects mean functions with their corresponding densities. Third, $\mathscr{H}(R, T)$ always exist and enjoys these basic properties, independent of the nature of T or the form of R.

Example. For the standard Wiener process on $T = [0, b]$, with covariance $R(s, t) = \min(s, t)$, the RKHS $\mathscr{H} = \mathscr{H}(R, T)$ has a very concrete form [Neveu (1968), pages 69–70 or Jørsboe (1968), pages 42–43]:

$$\mathscr{H} = \left\{m\colon m(t) = \int_0^t \gamma(u)\, du,\ \gamma \in L^2(T)\right\}.$$

To see this, let 1_E be the indicator function of the set E. Then we have the representation

$$R(s, t) = \int_0^b 1_{[0, s]}(u) 1_{[0, t]}(u)\, du.$$

It can be shown that the set $\{1_{[0, t]},\ t \in T\}$ spans $L^2(T)$, and that the map ψ: $R_t \to 1_{[0, t]}$ can be extended to an isomorphism $\psi\colon \mathscr{H} \to L^2(T)$ which is in fact an isometry by virtue of the fact that

$$\langle R_s, R_t\rangle = R(s, t) = \left(1_{[0, s]}, 1_{[0, t]}\right).$$

In particular, if $m \in \mathscr{H}$ and $\gamma = \psi(m) \in L^2(T)$, then

$$m(t) = \langle m, R_t \rangle = \left(\gamma, 1_{[0,t]}\right) = \int_0^t \gamma(u)\,du,$$

as claimed. Thus *$\mathscr{H}$ consists precisely of the primitives of the functions in $L^2(T)$ such that $m(0) = 0$, and ψ is seen to be differentiation.*

As an illustration, fix $c \in T$ and consider R_c. Since $\psi(R_c) = 1_{[0,c]}$, we have

$$\begin{aligned} R_c(s) = \int_0^s 1_{[0,c]}(u)\,du &= s, \quad \text{if } s < c \\ &\equiv c, \quad \text{if } s \geq c; \end{aligned}$$

this can also be seen directly from the fact that $R_c(s) = \min(s, c)$. Thus our model includes mean functions of the form

$$\begin{aligned} m(t) &= t, \quad \text{if } t < c \\ &\equiv c, \quad \text{if } t \geq c, \end{aligned}$$

for any constant $c \in [0, b]$, since $m = R_c$. But $\Lambda^{-1}(m) = \Lambda^{-1}(R_c) = X_c$, and so the corresponding measure Q has density

$$dQ/dP = e^{X_c}/E(e^{X_c}).$$

[It is a simple (but worthwhile) exercise to verify that this formula indeed gives $dP/dP = 1$.]

Returning to the general case, we now see that both H and $\mathscr{H}$ parametrize $\mathscr{P}$ in one-to-one fashion. The map $\mathscr{P} \to H$ is given by (2.1) or (2.2), while the correspondence $\mathscr{P} \to \mathscr{H}$ is given directly by

$$Q \to m_Q(\cdot) = E_Q(X). \tag{2.7}$$

Let $\{U_\alpha, \alpha \in A\}$ be a complete orthonormal (CON) basis (possibly uncountable) of H, and let $g_\alpha = \Lambda(U_\alpha)$, so that $\{g_\alpha, \alpha \in A\}$ is a CON basis of $\mathscr{H}$. Then every $Y \in H$ has a Fourier expansion

$$Y = \sum a_\alpha U_\alpha, \tag{2.8a}$$

and if $m = \Lambda(Y)$, then

$$m = \sum a_\alpha g_\alpha \tag{2.8b}$$

in $\mathscr{H}$. From elementary Hilbert space theory we know that there are at most countably many nonzero coefficients a_α in (2.8), that the sequence $\{a_\alpha\}$ is unique to Y (and to m), and that $\Sigma a_\alpha^2 < \infty$. Moreover, the correspondence,

$$Y \to \{a_\alpha\}, \qquad a_\alpha = (U_\alpha, Y), \tag{2.9}$$

is an isometric isomorphism between H and $\ell^2 = \ell^2(A)$; similarly,

$$m \to \{a_\alpha\} \tag{2.10}$$

defines an isometric isomorphism between $\mathscr{H}$ and ℓ^2.

We may summarize these correspondences in this diagram:

$$\begin{tikzcd}
& \mathscr{P} \arrow[dl, "(2.1)\text{ or }(2.2)"'] \arrow[dr, "(2.7)"] & \\
H \arrow[rr, "\Lambda"] \arrow[dr, "(2.9)"'] & & \mathscr{H} \arrow[dl, "(2.10)"] \\
& \ell^2 &
\end{tikzcd} \tag{2.11}$$

We will refer to $\mathscr{H}$ as the *original parameter space* of the model. Roughly speaking, densities live in H, mean functions live in $\mathscr{H}$, and we will do our estimation in ℓ^2. In fact, we will estimate the mean function m by estimating $\mathbf{a} \in \ell^2$ and writing

$$\hat{m} = \sum \hat{a}_\alpha g_\alpha,$$

using (2.8).

Example (*continued*). There are many ways to construct a CON basis of $\mathscr{H}$. Perhaps the most useful is the following. Let T be any compact interval and R any continuous covariance function on $T \times T$, and define the integral operator with kernel R in the usual way:

$$Rf(s) = \int_T R(s,t)f(t)\,dt$$

(we use R for the operator as well as the kernel). Then R is an operator on L^2 with a countable system of eigenfunctions φ_k and eigenvalues λ_k, $\lambda_k \geq 0$. By Proposition 3.11 of Neveu (1968), a CON basis of $\mathscr{H}(R,T)$ is given by the functions $g_k = \sqrt{\lambda_k}\varphi_k$, $\lambda_k > 0$.

To construct this system when $R(s,t) = \min(s,t)$ and $T = [0, b]$, we note that if $R\varphi = \lambda\varphi$ for some λ (necessarily positive), then φ satisfies the differential equation

$$\varphi'' + \lambda^{-1}\varphi = 0,$$

with boundary conditions

$$\varphi(0) = \varphi'(b) = 0.$$

Writing $\beta^2 = \lambda^{-1}$, we find that the normalized solutions are given by

$$\varphi_k(t) = (2/b)^{1/2}\sin\beta_k t, \qquad k \in \mathbb{Z}^+,$$

with eigenvalues $\lambda_k = \beta_k^{-2}$, where

$$\beta_k = \left(k - \tfrac{1}{2}\right)n/b.$$

Thus the functions

$$g_k(t) = \sqrt{\lambda_k}\varphi_k(t) = \beta_k^{-1}(2/b)^{1/2}\sin\beta_k t, \qquad k \in \mathbb{Z}^+,$$

are a CON basis for $\mathscr{H}$.

This basis has the very useful property that $U_k = \Lambda_k^{-1}(g_k)$ can be found explicitly (see Remark 2.2). In fact, using Proposition 3.9 of Neveu we see that in general

$$U_k = \lambda_k^{-1/2}\int_0^b X_t \varphi_k(t)\,dt,$$

a version of which is gotten by integrating $X_t(\omega)\varphi_k(t)$ at each $\omega \in \Omega$ [see Beder (1981), pages 31–33]. These random variables are automatically a CON basis for H since they correspond to the set $\{g_k\}$ under the Loève map. In our case they are given by

$$U_k = \beta_k(2/b)^{1/2}\int_0^b X_t \sin\beta_k t\,dt.$$

Since $g_k = \Lambda U_k$, we can now use (2.8) to evaluate Λ or Λ^{-1} explicitly, as desired. As a special application, consider the function $R_t \in \mathscr{H}$. Its expansion is

$$R_t = \sum a_k g_k,$$

where the Fourier coefficients can be gotten from the reproducing property:

$$a_k = \langle R_t, g_k\rangle = g_k(t).$$

But $\Lambda^{-1}(R_t) = X_t$, so that we have

$$X_t = \sum g_k(t) U_k.$$

This is the well-known *Karhunen–Loève expansion* of X_t. Grenander (1950) refers to the variables U_k as the *observable coordinates* of the process.

As noted above, there are many ways to construct a CON basis of $\mathscr{H}$ and corresponding bases of H. For instance, any CON basis of $L^2(T)$ will give rise (via ψ^{-1}) to a CON basis of $\mathscr{H}$ and hence (via Λ^{-1}) to a CON basis of H. In defining both Grenander's sieve and our new one for an arbitrary covariance, we will allow the CON bases of H and $\mathscr{H}$ to be arbitrary but fixed as long as they correspond to each other via Λ.

CONSEQUENCE 3: THE LIKELIHOOD FUNCTION AND RANDOM SAMPLES. From Lemma 2.1, it is easy to see that $E(e^Y) = \exp(2^{-1}\|Y\|^2)$. Thus the density (2.2) is

$$\frac{dQ}{dP} = \exp\left(Y - \frac{1}{2}\|Y\|^2\right). \tag{2.12}$$

EXAMPLE (*continued*). Let $m = R_c$ for some $c \in [0, b]$; then $m = \Lambda(X_c)$, and so m is the mean function corresponding to a measure Q whose density is

$$\begin{aligned} dQ/dP &= e^{X_c}/E(e^{X_c}) \\ &= \exp\left(X_c - \tfrac{1}{2}\|X_c\|^2\right) \\ &= \exp\left(X_c - \tfrac{1}{2}R(c,c)\right), \end{aligned}$$

since $\|X_c\|^2 = \mathrm{Var}_P(X_c)$.

In general, we will want to form the likelihood based on a sample of n complete observations of the process, say

$$\{X_t(\omega_1),\ t \in T\}, \ldots, \{X_t(\omega_n),\ t \in T\}.$$

This gives us n independent observations $Y(\omega_1), \ldots, Y(\omega_n)$ on any $Y \in H$. We may define these on a common sample space $(\Omega^n, \mathscr{A}^{n\otimes})$ in the usual way, letting $\omega = (\omega_1, \ldots, \omega_n) \in \Omega^n$ and defining $Y_1, \ldots, Y_n$ by

$$Y_i(\omega) = Y(\omega_i), \qquad i = 1, \ldots, n. \tag{2.13}$$

LEMMA 2.2. *For each $Q \in \mathscr{P}$, let $Q^{n\otimes}$ denote the corresponding product measure on $(\Omega^n, \mathscr{A}^{n\otimes})$, and let $\mathscr{P}^{(n)} = \{Q^{n\otimes}, Q \in \mathscr{P}\}$.*

(i) *If $Y_1, \ldots, Y_n$ are defined from $Y \in H$ by* (2.13), *then under $Q^{n\otimes}$ the Y_i are i.i.d. random variables, each with the same distribution as Y has under Q.*

(ii) *$\mathscr{P}^{(n)}$ satisfies assumptions* A_1–A_4 *with respect to $\{X_{ti},\ t \in T,\ i = 1, \ldots, n\}$.*

(iii) *If $Q \leftrightarrow Y$ via* (2.2), *then the density w.r.t. $P^{n\otimes}$ is*

$$\frac{dQ^{n\otimes}}{dP^{n\otimes}} = \exp\left[\sum_{i=1}^{n} Y_i - \frac{n\|Y\|^2}{2}\right]. \tag{2.14}$$

Our interest is in maximizing the likelihood, that is, in maximizing (2.14) over $Y \in H$ for each $\omega \in \Omega^n$. It is more convenient to re-express (2.14) first in terms of an ℓ^2-parametrization. Thus, let $\{U_\alpha,\ \alpha \in A\}$ be a CON basis of H, and let $Q \in \mathscr{P}$, $Y \in H$ and $\mathbf{a} \in \ell^2$ correspond via (2.11). For each $\alpha \in A$, we define $U_{\alpha 1}, \ldots, U_{\alpha n}$ by (2.13), and their sample mean by

$$\overline{U}_\alpha = \frac{1}{n}\sum_i U_{\alpha i}. \tag{2.15}$$

EXAMPLE (*continued*). For a sample of size n, let the ith observation on X_t be X_{ti}, $i = 1, \ldots, n$. Then for the standard Wiener process we have

$$\overline{U}_k = \frac{1}{n}\left(\frac{2}{b}\right)^{1/2} \beta_k \sum_{i=1}^{n} \int_0^b X_{ti} \sin \beta_k t\, dt.$$

REMARK 2.3. $\overline{U}_\alpha$ should also be indexed by n, but we will suppress the n to avoid complicating the notation. However, the dependence of $\overline{U}_\alpha$ on n should be kept in mind (see Remark 5.1).

LEMMA 2.3. *With Q, Y, $\mathbf{a}$ and $\{\overline{U}_\alpha,\ \alpha \in A\}$ as above, we have*

$$\frac{dQ^{n\otimes}}{dP^{n\otimes}} = \exp\left(n\left(\sum_\alpha a_\alpha \overline{U}_\alpha - \frac{1}{2}\|\mathbf{a}\|^2\right)\right). \tag{2.16}$$

PROOF. For each i, $Y_i = \Sigma_\alpha a_\alpha U_{\alpha i}$, so $\Sigma_i Y_i = n \Sigma_\alpha a_\alpha \overline{U}_\alpha$. Since (2.9) is an isometry, we also have $\|Y\| = \|\mathbf{a}\|$. The lemma now follows from (2.14). □

REMARK 2.4. A measure μ is said to *dominate* $\mathscr{P}$ if every $Q \in \mathscr{P}$ is absolutely continuous with respect to μ [Halmos and Savage (1949)]. In the present case there is no "natural" dominating measure, such as Lebesgue measure, either on $(\Omega, \mathscr{A})$ or on the space of sample paths. On the other hand, every measure P' in $\mathscr{P}$ dominates $\mathscr{P}$, so that one can form densities dQ/dP'; we have chosen P as the dominating measure only for convenience. While it is therefore incorrect to speak of "the" likelihood function, the equation

$$\frac{dQ}{dP} = \frac{dQ}{dP'}\frac{dP'}{dP},$$

shows that the method of maximum likelihood is not affected by the choice of dominating measure.

CONSEQUENCE 4: OBSERVABLES AND THEIR DISTRIBUTION. What is observable under $Q \in \mathscr{P}$ is not a random variable but merely the class of random variables which differ from it on a set of Q-measure zero. Since $\mathscr{P}$ is homogeneous, these classes are the same for all $Q \in \mathscr{P}$. In fact, the following is true. Recall that two norms are equivalent if convergence in one implies convergence in the other.

LEMMA 2.4. *For $Q \in \mathscr{P}$, $H_Q = H$ as sets, and the norms defined by $\| \ \|_Q$ and $\| \ \|$ are equivalent.*

PROOF. To show the norms are equivalent, let $\|X_n - X\| \to 0$. Then $X_n - X \to 0$ in P-probability:

$$\forall \varepsilon > 0, \qquad \lim_n P(|X_n - X| > \varepsilon) = 0.$$

But then

$$\forall \varepsilon > 0, \qquad \lim_n Q(|X_n - X| > \varepsilon) = 0,$$

as $Q \sim P$. Since X and X_n are normal under Q, this implies $\|X_n - X\|_Q \to 0$ [Neveu (1968), Lemma 1.5]. □

From this it is not hard to extend (2.3) as follows:

COROLLARY 2.2. *If $Q \in \mathscr{P}$, then* $\mathrm{Cov}_Q(X, Y) = \mathrm{Cov}(X, Y)$ *for all $X, Y \in H$.*

We will be particularly interested in the following question. Let $\{U_\alpha, \alpha \in A\}$ be a CON basis of H. If we consider the U_α as random variables, what is their distribution under each measure in $\mathscr{P}$? We see from Lemma 2.1 that under P they are i.i.d. $N(0, 1)$. In general we have the following:

COROLLARY 2.3. *Let $\{U_\alpha, \alpha \in A\}$ be a CON set in H. Let $Q \in \mathscr{P}$ correspond to $\mathbf{a} = \{a_\alpha\} \in \ell^2$ via* (2.11). *Then under Q the U_α are independent, and U_α has the $N(a_\alpha, 1)$ distribution. If we define $\overline{U}_\alpha$ by* (2.15) *for a sample of size n, then under $Q^{n\otimes}$ the $\overline{U}_\alpha$ are independent, and $\overline{U}_\alpha$ has the $N(a_\alpha, 1/n)$ distribution.*

PROOF. We need only note that if Y corresponds to Q and $\mathbf{a}$ via (2.11), then

$$E_Q(U_\alpha) = (U_\alpha, Y) = a_\alpha. \square$$

EXAMPLE (*continued*). Consider the probability Q corresponding to the mean function $m = R_c$ for some $c \in [0, b]$. Let $\{U_k\}$ be the orthonormal basis constructed earlier. Now $X_c = \Lambda^{-1}(m)$ has the Karhunen–Loève expansion

$$X_c = \sum g_k(c) U_k,$$

and so we immediately "read off" the distribution of U_k under Q as being $N(g_k(c), 1)$; for a sample of size n, $\bar{U}_k$ has the $N(g_k(c), 1/n)$ distribution under $Q^{n\otimes}$.

REMARK 2.5. It may be worth pausing at this point to relate the results of this section to the familiar case $T = \{1, \dots, p\}$ of multivariate analysis. Now the quantity $R(s, t)$ is the (s, t)th entry of the covariance matrix Σ of the vector $\mathbf{X} = (X_1, \dots, X_p)'$; we will assume Σ is invertible for simplicity.

Let Q be the measure which gives $\mathbf{X}$ the $N(\mu, \Sigma)$ distribution, and P, the $N(\mathbf{0}, \Sigma)$ distribution. Then the induced measures $\tilde{Q}$ and $\tilde{P}$ on $\mathbb{R}^n$ have likelihood ratio

$$(2.17) \quad d\tilde{Q}/d\tilde{P} = f(\mathbf{x}|\mu, \Sigma)/f(\mathbf{x}|\mathbf{0}, \Sigma) = \exp\left(\mu'\Sigma^{-1}\mathbf{x} - (1/2)\mu'\Sigma^{-1}\mu\right),$$

as is easily seen. It follows that on $(\Omega, \mathscr{A})$ we have

$$(2.18) \qquad dQ/dP = \exp\left(\mu'\Sigma^{-1}\mathbf{X} - (1/2)\mu'\Sigma^{-1}\mu\right),$$

which is precisely of form (2.12) with

$$Y = \mu'\Sigma^{-1}\mathbf{X}.$$

H is now simply the vector space of linear combinations of $X_1, \dots, X_p$, while $\mathscr{H}$ is the set $\mathbb{R}^n$ equipped with inner product $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}'\Sigma^{-1}\mathbf{b}$, and we have $\Lambda(Y) = \mu$.

As we remarked earlier, it is not necessary in this case to form the likelihood dQ/dP, since we can work with the induced density $f(\mathbf{x}|\mu, \Sigma) = d\tilde{Q}/d$ (Lebesgue measure). Using (2.17) or just $f(\mathbf{x}|\mu, \Sigma)$, the MLE $\hat{\mu}$ is the observed value $\mathbf{x}$, which is $\mathbf{X}(\omega)$ for some $\omega \in \Omega$. Thus to use (2.18) as a likelihood, we fix $\omega \in \Omega$ and so fix $\mathbf{X}(\omega)$; for each ω, then, (2.18) becomes a function of μ, which we can maximize.

If we write (2.18) in the form (2.12), then for each $\omega \in \Omega$ we would like to fix $Y(\omega)$ and maximize over μ, but $Y(\omega) = \mu'\Sigma^{-1}\mathbf{X}(\omega)$ itself depends on μ as well. In the case that T is an interval or some other set, the dependence of Y on $(X_t, t \in T)$ is much less evident, and so the maximization problem is, on the face of it, much more difficult. By fixing a CON basis (U_α) of H we reestablish a coördinate system in which to represent Y and therefore the density dQ/dP. While the variables X_t no longer appear directly in the density, they still determine it through the variables U_α. To form the likelihood we still fix $\omega \in \Omega$, thereby fixing $\{X_t(\omega), t \in T\}$ and so fixing $\{U_\alpha(\omega), \alpha \in A\}$ as well. The density is now a function solely of the parameter. The same applies to a sample of size n, and maximizing this density is the problem we consider next.

We must exercise some caution in taking this maximum. Each density is defined only up to sets of measure zero, while the likelihood function, which evaluates each density at a point ω, requires us to choose a version of each density first. Let us agree to say that the likelihood L is almost surely unbounded if for every version $\tilde{L}$ of L there is an event E of probability 1 such that $\tilde{L}$ is unbounded for every $\omega \in E$. Except in certain cases (such as when $\dim H < \infty$), it will not be possible to choose a single event E which works for all versions of L.

3. MLE's and a 0-or-1 property. Give a sample of n complete observations of the process, we wish to maximize (2.16) by fixing $\omega \in \Omega^n$ and allowing $\mathbf{a}$ to vary over the parameter space ℓ^2. It will be useful to solve a slightly more general problem of maximizing over (closed) subspaces of the parameter space.

Thus let $B \subset A$, and consider the subspace $H_0 \subset H$ spanned by $\{U_\alpha, \alpha \in B\}$, corresponding directly to a subspace $\ell^2(B)$ of $\ell^2(A)$. The following proof is essentially due to Rozanov [(1971), page 128].

THEOREM 3.1. *Let $\ell^2(B)$ be defined as above. If B is finite, then the likelihood* (2.16) *is maximized over $\ell^2(B)$ at $\hat{\mathbf{a}} = \{\hat{a}_\alpha\}$ given by*

$$\hat{a}_\alpha = \overline{U}_\alpha, \quad \text{if } \alpha \in B, \qquad = 0, \quad \text{otherwise.} \tag{3.1}$$

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

PROOF. If $\mathbf{a} \in \ell^2(B)$, then $a_\alpha = 0$ for $\alpha \notin B$, so the sum in (2.16) is over $\alpha \in B$.

If B is finite, then we write (2.16) as

$$\frac{2}{n}\ln\frac{dQ^{n\otimes}}{dP^{n\otimes}} = \sum_{\alpha \in B}\left(2a_\alpha\overline{U}_\alpha - a_\alpha^2\right).$$

For each fixed value $\overline{U} = \overline{U}(\omega)$, this is maximized when each term is maximized, namely at $\hat{a}_\alpha = \overline{U}_\alpha$.

If B is infinite, then by considering the subspace $\ell^2(B_0) \subset \ell^2(B)$ for a countable set $B_0 \subset B$ we see that we may assume that B itself is countable. In this case, by considering (2.16) with the "basis" elements $\mathbf{a} = (0 \cdots 010 \cdots)$ of $\ell^2(B)$ and the corresponding measures $Q = P_{\mathbf{a}} \in \mathscr{P}$, we have

$$n^{-1}\sup\ln\frac{dQ^{n\otimes}}{dP^{n\otimes}} = \sup_\alpha\left\{\overline{U}_\alpha, \alpha \in B\right\} - \frac{1}{2};$$

but by Corollary 2.3 the latter is infinite a.s. ($P^{n\otimes}$), and thus a.s. ($\mathscr{P}^{(n)}$). Thus, for a fixed version of the likelihood there is an event $E \in \mathscr{A}^{n\otimes}$ of probability 1 such that the likelihood is unbounded on a countable subset of $\ell^2(B)$, and so a fortiori on all of $\ell^2(B)$, for every $\omega \in E$. □

In terms of the original parameter space $\mathscr{H}$, we can give the following "coordinate-free" version of the theorem:

COROLLARY 3.1. *Let $\mathscr{H}_0$ be a closed subspace of $\mathscr{H}$. If* $\dim \mathscr{H}_0 < \infty$, *then the likelihood function L can always be maximized over $\mathscr{H}_0$. If* $\dim \mathscr{H}_0 = \infty$, *$L$ is almost surely unbounded over $\mathscr{H}_0$. In particular, if* $\dim \mathscr{H} = \infty$, *then the MLE of the mean almost surely fails to exist.*

Moreover, when $\dim \mathscr{H}_0 = d < \infty$, *let $\{g_1, \dots, g_d\}$ be an orthonormal basis of $\mathscr{H}_0$, and let $U_k = \Lambda^{-1} g_k \in H$. Then the likelihood is maximized over $\mathscr{H}_0$ at*

$$m = \sum \overline{U}_k g_k, \tag{3.2}$$

where $\overline{U}_k$ is defined by (2.15).

This result was given by Parzen [(1961a), page 979 and (1961b), page 482] under some assumptions about the index set T. In the latter paper he showed that the "formal" solution given by (3.1) in the infinite-dimensional case yields a sequence $\hat{\mathbf{a}}$ which is almost surely not in ℓ^2. This is not quite a complete proof of the failure of maximum likelihood, but one can use it to get an alternate proof of Theorem 3.1; see Beder (1982).

EXAMPLE (*continued*). For the Wiener process, $\mathscr{H}$ has countably infinite dimension, so that the likelihood function for the mean is unbounded almost surely.

REMARK 3.1. The existence of the MLE for the mean is an event of probability 0 or 1, depending only on whether $\dim(\mathscr{H})$ is infinite or finite. It is natural to call this dimension the *dimension of the process* (with respect to the model $\mathscr{P}$). Note that it equals $\dim(H)$, and so by Corollary 2.2 it is the "degrees of freedom" of the process (w.r.t. $\mathscr{P}$), i.e., the number of uncorrelated random variables which may be formed from the set $\{X_t,\ t \in T\}$ by taking linear combinations and mean-square limits.

REMARK 3.2. If the functions g_k in (3.2) are chosen merely to be linearly independent, it is still possible to give a simple description of the MLE. For this and a discussion of some likelihood ratio tests, see Beder [(1982), pages 11–12].

Theorem 3.1 and the corollary deal only with maximizing the likelihood over subspaces. One can seek the maximum over other sets in the parameter space, and that is in fact how we will generalize the GGH sieve estimator.

4. The sieve estimator. The unboundedness of the likelihood function for *any* sample size n leads us to consider sieve estimation [Grenander (1981), page 357]. 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 $\{\mathscr{S}_d\}$ of subsets of Θ indexed by a parameter d such that

(a) $d' > d \Rightarrow \mathscr{S}_{d'} \supset \mathscr{S}_d$,

(b) $\cup \mathscr{S}_d$ is dense in Θ, and

(c) the likelihood can be maximized over each $\mathscr{S}_d$ (for all sample sizes n, or at least for n sufficiently large).

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

Grenander originally parametrized the sieve by the index $\mu = 1/d$, called the "mesh size;" in this case one must alter (a) accordingly. In either case the sieve index is usually taken to be real, although one could certainly index a sieve by an element of a directed set, for example.

The denseness of $\cup\mathscr{S}_d$ in Θ presumes a topology, which is natural as one is ultimately interested in some sort of consistency result $\hat{\theta} \to \theta$ as d and $n \to \infty$.

The GGH sieve is defined as follows. Assume that $\mathscr{H}$ is separable and infinite-dimensional (see Remark 4.1 below), and fix a CON basis $\{g_k,\ k \in \mathbb{Z}^+\}$ corresponding to a basis $\{U_k,\ k \in \mathbb{Z}^+\}$ of H. Reparametrize by $\ell^2(\mathbb{Z}^+)$ in the usual way [see (2.11)]. Then the sets

$$\mathscr{S}_d = \left\{\mathbf{a} \in \ell^2 \colon \sum k^2 a_k^2 \le d\right\}, \qquad d > 0, \tag{4.1}$$

form a sieve in ℓ^2, where convergence in ℓ^2 is defined by the norm as usual. For a sample of size n, the resulting sieve estimator $\hat{\mathbf{a}}_{dn}$ is given by

$$\hat{a}_{dnk} = \frac{\overline{U}_k}{1 + \lambda k^2},$$

where $\lambda = \lambda_{dn}$ satisfies

$$\sum \frac{k^2 \overline{U}_k^2}{(1 + \lambda k^2)^2} = d \tag{4.2}$$

and $\overline{U}_k$ is defined by (2.15). Geman and Hwang (1982) show that if $d = d_n$ is chosen to depend on the sample size in such a way that $d_n \to \infty$ and $d_n = O(n^{1/3-\varepsilon})$ for some $\varepsilon > 0$, then $\|\hat{\mathbf{a}}_{dn} - \mathbf{a}\| \to 0$ a.s. $P_{\mathbf{a}}$ ($P_{\mathbf{a}}$ = the measure in $\mathscr{P}$ corresponding to $\mathbf{a}$).

The difficulty in using this estimator is that it depends on the quantity λ, a random variable defined implicitly from the data by (4.2). This not only makes its value difficult to compute, but also probably precludes the possibility of making small sample statements about the behavior of $\hat{\mathbf{a}}$.

Let us instead define

$$\mathscr{S}_d = \left\{\mathbf{a} \in \ell^2 \colon a_k = 0 \text{ for } k > d\right\}, \qquad d = 1,2,3,\dots. \tag{4.3}$$

It follows from Theorem 3.1 that $\{\mathscr{S}_d,\ d \in \mathbb{Z}^+\}$ is a sieve, where the resulting sieve estimator $\hat{\mathbf{a}}_{dn}$ is given by

$$\hat{\mathbf{a}}_{dn} = \left(\overline{U}_1, \dots, \overline{U}_d, 0, \dots\right) \tag{4.4}$$

based on a sample of size n. This corresponds to the sieve $\{\mathscr{H}_d,\ m \in \mathbb{Z}^+\}$ in $\mathscr{H}$, where $\mathscr{H}_d$ is the span of $\{g_1, \dots, g_d\}$. The sieve estimator of the mean function is thus given exactly by

$$\hat{m}(t) = \sum_{k=1}^{d} \overline{U}_k g_k(t). \tag{4.5}$$

If we consider $\{\hat{m}(t),\ t \in T\}$ as a stochastic process, then Corollary 2.3 easily gives its distribution:

THEOREM 4.1. *Under $P_{\mathbf{a}}$ the sieve estimator $\hat{m}(t)$ defined by* (4.5) *is a Gaussian process with mean*

$$\sum_{k=1}^{d} a_k g_k(t) \tag{4.6}$$

and covariance

$$n^{-1} \sum_{k=1}^{d} g_k(s) g_k(t). \tag{4.7}$$

From Theorem 4.1 we can quickly derive several "local" results:

COROLLARY 4.1. *Suppose that $\dim \mathscr{H}$ is countably infinite. If $d \to \infty$, then at each $t \in T$*

(i) *$\hat{m}(t)$ is asymptotically unbiased for $m(t)$, and*

(ii) *$\hat{m}(t)$ converges in probability and in mean square to $m(t)$ if in addition $d = O(n)$ as $n \to \infty$.*

PROOF. Asymptotic unbiasedness is obvious from (4.6). Since mean-square error = variance + (bias)^2, we need to show that $\mathrm{Var}(\hat{m}(t)) \to 0$ if $d = O(n)$. Now the variance under $P_{\mathbf{a}}$ is given by (4.7), with $s = t$. But in *any* reproducing kernel Hilbert space $\mathscr{H}(R, T)$ with *any* CON basis $\{g_\alpha,\ \alpha \in A\}$, the expansion,

$$R(s,t) = \sum_{\alpha} g_\alpha(s) g_\alpha(t),$$

holds for all $s, t \in T$ [see, e.g., Halmos (1967), Problem 30]. Thus in particular the infinite sum $\Sigma(g_k(t))^2$ converges at each $t \in T$, and so an application of the Toeplitz lemma [Loève (1977)] shows that $\mathrm{Var}(\hat{m}(t)) \to 0$ as $d \to \infty$ and $n \to \infty$ as long as $d = O(n)$.

Finally, mean-square convergence and an application of Chebyshev's inequality show that $\hat{m}(t) \to m(t)$ in probability. □

REMARK 4.1. Both the GGH sieve and the new one satisfy property (b) of Definition 4.1 as long as $\mathscr{H}$ is separable (in which case, in fact, $\cup \mathscr{S}_m$ is not just dense but actually equals ℓ^2). Corollary 4.1 also depends upon this assumption. We may broaden the applicability of both sieves by making a slightly more general assumption, namely:

(S) The true mean is an element of a known separable subspace $\mathscr{H}_0$ of $\mathscr{H}$.

In this case we take $\{g_k, k \in \mathbb{Z}^+\}$ to be a CON basis of $\mathscr{H}_0$, corresponding to a CON basis $\{U_k, k \in \mathbb{Z}^+\}$ of a separable subspace H_0 of H. We will discuss assumption (S) in Section 6.

5. Consistency. Our goal is to show that the new sieve estimator (4.4) is "globally" strongly consistent for the true parameter $\mathbf{a}$ if $d = d_n$ is chosen to go to infinity at an appropriate rate. Convergence of $\hat{\mathbf{a}}$ to $\mathbf{a}$ is in ℓ^2: we want to show that the square error $\|\hat{\mathbf{a}} - \mathbf{a}\|^2$ goes to 0 a.s. $(P_{\mathbf{a}})$. As we noted in the introduction, if $m \leftrightarrow \mathbf{a}$ and $\hat{m} \leftrightarrow \hat{\mathbf{a}}$ via (2.8), then $\|\hat{m} - m\| = \|\hat{\mathbf{a}} - \mathbf{a}\|$ (the first norm in $\mathscr{H}$), so that we will have consistency in the original parameter space as well.

Fix $\mathbf{a} \in \ell^2$, and let the corresponding measure in $\mathscr{P}$ be $P_{\mathbf{a}}$. For the sieve estimator $\hat{\mathbf{a}} = \hat{\mathbf{a}}_{dn}$ given by (4.4), we have

$$\begin{aligned} \|\hat{\mathbf{a}} - \mathbf{a}\|^2 &= \sum_{k \le d} \left(\overline{U}_k - a_k\right)^2 + \sum_{k > d} a_k^2 \\ &= Z_{dn} + \sum_{k>d} a_k^2, \end{aligned} \tag{5.1}$$

say. Now the second term, which is nonstochastic, goes to zero as long as $d \to \infty$. We thus need to choose $d = d_n$ so that $Z_{dn} \to 0$ a.s. $(P_{\mathbf{a}})$. At this point we can get weak convergence fairly simply.

LEMMA 5.1. *Under $P_{\mathbf{a}}$, nZ_{dn} has the $\chi^2(d)$ distribution. If $d/n \to \beta < \infty$, then $Z_{dn} \to \beta$ in $P_{\mathbf{a}}$-probability.*

PROOF. From Corollary 2.3 we see that under $P_{\mathbf{a}}$ (i.e., under $P_{\mathbf{a}}^{n\otimes}$) the variables $\sqrt{n}(\overline{U}_k - a_k)$ are i.i.d. $N(0,1)$, so that nZ_{dn} is $\chi^2(d)$. It is easy to show then that the mgf (moment generating function) of Z_{dn} converges to $\exp(\beta t)$, which is the mgf of the degenerate distribution at β. □

REMARK 5.1. As we noted earlier (Remark 2.3), $\overline{U}_k$ depends on both k and n, so that Z_{dn} is the sum of a row in a triangular array. Therefore, we cannot infer the strong convergence of Z_{dn} from the law of large numbers in the usual way.

Lemma 5.1 tells us what we can expect at most from strong consistency. We must have $d = o(n)$, and if $d/n \to \beta > 0$, then $\hat{\mathbf{a}}$ will not even be weakly consistent.

Since $Z_{dn} > 0$ for all d and n, $Z_{dn} \to 0$ as n and $d \to \infty$ iff for every $\varepsilon > 0$,

$$P_{\mathbf{a}}(Z_{dn} > \varepsilon \text{ i.o.}) = 0. \tag{5.2}$$

The Borel–Cantelli lemma gives a sufficient condition for (5.2): for every $\varepsilon > 0$,

$$\sum_n P_{\mathbf{a}}(Z_{dn} > \varepsilon) < \infty. \tag{5.3}$$

Thus we must choose $d = d_n$ so that (5.3) holds for every $\varepsilon > 0$. Let us start by getting a convenient upper bound for each term in the sum.

LEMMA 5.2. *$P_{\mathbf{a}}(Z_{dn} > \varepsilon) < [(d + 2(k - 1))/n\varepsilon]^k$ for any $k \in \mathbb{Z}^+$.*

PROOF. Since $Z_{dn} > 0$, Chebyshev's inequality gives

$$P_{\mathbf{a}}(Z_{dn} > \varepsilon) = P_{\mathbf{a}}\left[(nZ_{dn})^k > (n\varepsilon)^k\right] \leq \frac{E\left((nZ_{dn})^k\right)}{(n\varepsilon)^k},$$

for all $k \in \mathbb{Z}^+$. But nZ_{dn} has the $\chi^2(d)$ distribution, so

$$E\left((nZ_{nd})^k\right) = d(d+2)\cdots(d+2(k-1)) < (d+2(k-1))^k. \qquad \square$$

The bound

$$\left[\frac{d + 2(k-1)}{n\varepsilon}\right]^k \tag{5.4}$$

in Lemma 5.2 introduces a new "parameter", k, which we will allow to depend on n. Thus we have two sequences to pick, $\{d_n\}$ and $\{k_n\}$, in order to make (5.4) a summable sequence. This extra latitude is just what we need.

Our main aim is to pick $\{k_n\}$ so that (5.4) is summable when $d \to \infty$ and $d/n \to 0$, if possible, but by solving this problem more generally when $d/n \to \beta \geq 0$ we can get a bit more insight into the behavior of Z_{dn} when it fails to converge to zero. The following is an easy application of the Cauchy root test.

LEMMA 5.3. *Let $d/n \to \beta \geq 0$, and let $k/n \to \gamma > 0$ as $n \to \infty$. If $\beta + 2\gamma < \varepsilon$, then (5.4) is summable.*

THEOREM 5.1. *If $d = d_n$ is such that $d \to \infty$ and $d/n \to \beta \geq 0$ as $n \to \infty$, then $P_{\mathbf{a}}(Z_{dn} > \varepsilon$ i.o.$) = 0$ for every $\varepsilon > \beta$.*

PROOF. Fix $\varepsilon > \beta$, and choose $\{k_n\}$ such that

$$\frac{k}{n} \to \gamma \in \left(0, \frac{\varepsilon - \beta}{2}\right).$$

Then (5.4) is summable by Lemma 5.3, and so the conclusion follows. □

When $\beta = 0$ we get $P_{\mathbf{a}}(Z_{dn} > \varepsilon$ i.o.) for every $\varepsilon > 0$, which implies our main result:

COROLLARY 5.1. *If $d \to \infty$ and $d/n \to 0$, then $\|\hat{\mathbf{a}}_{dn} - \mathbf{a}\| \to 0$ a.s. $P_{\mathbf{a}}$.*

We have noted that this is in some sense best possible, as we fail to get even weak consistency when $d/n \to \beta > 0$. Interestingly, Theorem 5.1 tells us what does happen in this case, too.

COROLLARY 5.2. *If $d/n \to \beta > 0$, then $Z_{dn} < \varepsilon$ a.s. $P_{\mathbf{a}}$ for every $\varepsilon > \beta$, and $\|\hat{\mathbf{a}}_{dn} - \mathbf{a}\|$ is bounded a.s. $P_{\mathbf{a}}$ (uniformly in n).*

REMARK 5.2. When $d/n \to \beta \geq 0$ we have $d = O(n)$, so that the results of Corollary 4.1 hold automatically.

6. Conclusion.

We have used sieve estimation to construct a point estimator of the mean of a Gaussian process of arbitrary known covariance. The estimator is given explicitly in terms of the data, has a simple distribution and is strongly consistent. Our construction applies to discrete- and continuous-time processes, to nonstationary processes, to random fields and to more general Gaussian processes. For all of these, the norm of the reproducing kernel Hilbert space $\mathscr{H}(R, T)$ provides the natural metric for convergence.

Assumption (S) (Section 4) is the only restriction we have imposed. It is needed both for the new sieve and for the one given by Grenander, and appears to be a natural limitation on the method of sieves. Clearly, without (S) both sieve estimators are strongly inconsistent when $\dim(\mathscr{H})$ is uncountable. To eliminate the assumption, we need some way of "pretesting" all separable subspaces $\mathscr{H}_0$ of $\mathscr{H}$ to find one which contains the true mean. It is not clear how to go about this.

We would also like to be able to construct confidence sets for the true mean function. What stands in the way at present is the bias of our estimator. In general, the method of sieves has been used to construct consistent point estimators; distribution theory and "rate of convergence" questions have received little attention. It may be hoped that Theorem 4.1 will give us a handle on this problem in the present case.

Acknowledgments. I would like to thank Jugal Ghorai, Ian McKeague, Anestis Antoniadis and the referee for important and helpful comments. This work is based in part on my dissertation, and I would like to thank my advisor, Robert H. Shumway, for his guidance and encouragement.

REFERENCES

ANTONIADIS, A. (1985). Parametric estimation for the mean of a Gaussian process by the method of sieves. Technical Report, Séminaire de Statistique, Université de Grenoble.

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

BEDER, J. H. (1981). Likelihood methods for Gaussian processes. Ph.D. dissertation, The George Washington Univ., Washington.

BEDER, J. H. (1982). Maximum likelihood estimation for Gaussian processes. Technical Report 35, Division of Statistics, Univ. California, Davis.

BEDER, J. H. (1986). A sieve estimator for the covariance of a Gaussian process. Submitted.

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

GRENANDER, U. (1950). Stochastic processes and statistical inference. *Ark. Mat.* **1** 195–276.

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

JØRSBOE, O. G. (1968). *Equivalence or Singularity of Gaussian Measures on Function Spaces.* Various Publications Series No. 4, Matematisk Institut, Aarhus Universitet, Denmark.

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. and OODAIRA, H. (1963). The equivalence and singularity of Gaussian measures. In *Proc. of the Symposium on Time Series Analysis* (M. Rosenblatt, ed.) 279–291. Wiley, 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. (1985). The method of sieves: a survey of recent applications. Statistics Report M-718, Florida State Univ., Tallahassee. (To appear under the entry "Sieves, method of" in the *Encyclopedia of Statistical Sciences*. Wiley, New York.)

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

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

NGUYEN, H. T. and PHAM, T. D. (1982). Identification of nonstationary diffusion model by the method of sieves. *SIAM J. Control Optim.* **20** 603–611.

PARZEN, E. (1959). Statistical inference on time series by Hilbert space methods, I. Technical Report 23, Statistics Dept., Stanford Univ. Reprinted in Parzen (1967), Paper 13.

PARZEN, E. (1961a). An approach to time series analysis. *Ann. Math. Statist.* **32** 951–989. Reprinted in Parzen (1967), Paper 14.

PARZEN, E. (1961b). Regression analysis of continuous parameter time series. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 469–489. Univ. California Press. Reprinted in Parzen (1967), Paper 15.

PARZEN, E. (1967). *Time Series Analysis Papers*. Holden-Day, San Francisco.

ROZANOV, JU. A. (1971). Infinite-dimensional Gaussian distributions. *Proc. of the Steklov Institute of Mathematics* **108**. Trans. G. Biriuk. Amer. Math. Soc., Providence, R.I.

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