INTRODUCTION TO GAUSSIAN PROCESSES

BY STEVEN P. LALLEY

Table of contents

1. Definitions and Examples

Definition 1. *A Gaussian process $\{X_t\}_{t\in T}$ indexed by a set T is a family of (real-valued) random variables X_t, all defined on the same probability space, such that for any finite subset $F\subseteq T$ the random vector $X_F=(X_{t_1},X_{t_2},\dots,X_{t_n})$ has a (possibly degenerate) Gaussian distribution; if these finite-dimensional distributions are all non-degenerate then the Gaussian process is said to be non-degenerate. Equivalently, $\{X_t\}_{t\in T}$ is Gaussian if every finite linear combination $\sum_{t\in F}a_t X_t$ is either identically zero or has a Gaussian distribution on $\mathbb{R}$. A Gaussian process $\{X_t\}_{t\in T}$ is centered if $E X_t=0$ for every $t\in T$, and its covariance function is the (symmetric) bivariate function*

$$R(s,t)=\operatorname{cov}(X_s,X_t)=E(X_s X_t)-E X_s\, E X_t \tag{1}$$

Definition 2. *The canonical metric d associated with a non-degenerate Gaussian process $\{X_t\}_{t\in T}$ is*

$$d(s,t)=\sqrt{E(X_s-X_t)^2} \tag{2}$$

Covariance Functions: Recall that the covariance matrix of a multivariate Gaussian random vector is a symmetric, nonnegative definite matrix; the distribution is said to be non-degenerate if its covariance matrix is strictly positive definite. The mean vector and covariance matrix uniquely determine a Gaussian distribution; consequently, the *mean function* and *covariance function* of a Gaussian process completely determine all of the finite-dimensional distributions (that is, the joint distributions of finite subsets X_F of the random variables). Thus, if two Gaussian processes $X = \{X_t\}_{t\in T}$ and $Y = \{Y_t\}_{t\in T}$ have the same mean and covariance functions, then for any event B

$$P(X \in B) = P(Y \in B).$$

Since any event can be arbitrarily well-approximated by events that depend on only finitely many coordinates, it follows that the equality holds for *all* events B. Therefore, the processes X and Y are identical in law.

2. Continuous Extensions and Maxima of Gaussian Processes

2.1. Continuity

Keep in mind that there are covariance functions R that exist such that no centered Gaussian process with covariance function R has continuous sample paths and there are also *continuous* covariance functions that exist that do not have versions with continous sample paths.

Definition 3. *Let (T,d) be a compact metric space. For each $\varepsilon > 0$ the Lebesgue covering number $N(\varepsilon)$ is the minimum number of (open) ε-balls needed to cover T.*

Theorem 4. *Let d be the canonical metric of a non-degenerate, centered Gaussian process $\{X_i\}_{i\in T}$, and let $N(\varepsilon)$ be the Lebesgue covering function. If for some $\rho > 0$*

$$\int_0^{\varepsilon} \log N(\eta)\, d\eta < \infty$$

then the Gaussian process has a version with uniformly continuous sample paths. Consequently (since T is compact), for this version,

$$\sup_{i\in T} X_i = \max_{i\in T} X_i \quad a.s.$$

Remark 5. This result can then also be used as an extension theorem. In particular, if (T,d) is not a complete metric space then the theorem implies that under the assumption $\int_0^{\varepsilon} \log N(\eta)\, d\eta < \infty$ the Gaussian process has a uniformly continuous extension to the completion of (T,d), by the next lemma.

Lemma 6. *Let D be a dense subset of a metric space T, and let $f: D \to \mathbb{R}$ be uniformly continuous on D, i.e., for each $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(y) - f(x)| < \varepsilon$ for any two points $x, y \in D$ at distance $d(x,y) < \delta$. Then there is a unique uniformly continuous extension of f to T.*

Proof. This is a routine exercise in elementary real analysis.

Proof of Theorem 1. Assume without loss of generality that $\rho = 1$ (if not, rescale). Fix $r \geq 4$, and for each $n = 0, 1, 2, \ldots$ let T_n be a subset of T with cardinality $N(r^{-n})$ such that the r^{-n}-balls centered at the points of T_n cover T. Clearly, the set $D = \bigcup_{n \geq 0} T_n$ is a countable dense subset of T. The plan will be to use Borel-Cantelli to show that with probability one the restriction of the Gaussian process to the index set D is uniformly continuous; Lemma 2.1 will then imply that there is a continuous extension to T.

For each $n \geq 0$ define $D_n = \bigcup_{k=0}^{n} T_k$; the sets D_n are nested, and D_n has cardinality $N_n \leq (n+1)\, N(r^{-n})$, and so the hypothesis (2.1) implies that

$$\sum_{n=0}^{\infty} r^{-n} A_n < \infty \quad \text{where} \quad A_n = \sqrt{\log\left(4^n N_n\right)}$$

Consider the events

$$B_n = \{\exists\, s, t \in D_n : |X_s - X_t| \geq A_n\, d(s,t)\}$$

Since there are at most N_n^2 pairs of points in D_n, Bonferroni implies that

$$P(B_n) \leq C\, N_n^2\, e^{(-n \log 4 - \log N_n^2)} \leq C\, 4^{-n}$$

(Here C is any constant such that if Z is a standard Gaussian random variable then $P(|Z| > x) \leq C \exp(-x^2/2)$ for all $x \geq 0$.) This is summable, so Borel-Cantelli implies that with probability one only finitely many of the events B_n occur. In particular, w.p.1 there exists $K < \infty$ (random) such that $\sum_{n \geq 1} B_n = 0$. We will show that this implies that the mapping $s \to X_s$ is uniformly continuous on D.

Fix $\varepsilon > 0$, and let $m = m(\varepsilon)$ be such that $\sum_{n \geq m} r^{-n} A_n < \varepsilon$. Suppose that $s, t \in D$ are any two points at distance less than r^{-m}. Let $l \geq 0$ be the smallest nonnegative integer such that $s, t \in D_{m+1}$. Then there is a chain of points

$$s = s_k \in T_{n_k}, s_{k-1} \in T_{n_{k-1}}, \ldots, s_0 \in T_n$$

$$t = t_l \in T_{n_l}, t_{l-1} \in T_{n_{l-1}}, \ldots, t_0 \in T_n$$

connecting s to t such that at every link of the chain

$$d(s_r, s_{r-1}) \leq 4^{-n+r+2}$$

$$d(t_r, t_{r-1}) \leq 4^{-n-r+2}$$

$$d(s_0, t_0) \leq 4^{-n+2}$$

If $m \geq K$ (as it will be eventually as $\varepsilon \to 0$ then along any link (u, v) of these chains $|X_u - X_v| < A_j\, d(u, v)$, where j is the depth of the link. It then follows that

$$|X_s - X_t| < 3\,\varepsilon \tag{3}$$

Thus, the restriction of the Gaussian process to D is, with probability one, uniformly continuous.

Corollary 7. *If T is a compact subset of $\mathbb{R}^k$ and if $R(s,t)$ is a symmetric, Lipschitz continuous, nonnegative definite kernel on $T \times T$, then there is a centered Gaussian process $\{X_t\}_{t\in T}$ with continuous sample paths and covariance function $R(s,t)$.*

Proof. Exercise.

Corollary 8. *Let T be a compact interval in $\mathbb{R}^1$ and assume that $R(s,t)$ satisfies the hypotheses of Corollary 2.2. If R is twice continuously differentiable, and if the mixed second partials are Lipschitz continuous on $T \times T$, then the centered Gaussian process $\{X_t\}_{t\in T}$ with covariance function $R(s,t)$ has a differentiable version, and the derivative process $\{X_t'\}_{t\in T}$ is a centered Gaussian process with covariance function*

$$\operatorname{cov}(X_s', X_t') = \frac{\partial^2 R}{\partial s\, \partial t}(s,t) \tag{4}$$

Proof (Sketch). Define a Gaussian process on $T \times T \setminus \{(t,t)\}_{t\in T}$ by

$$Y_{s,t} = \frac{X_t - X_s}{t - s} \tag{5}$$

The hypotheses on R guarantee that the covariance function of the process $\{Y_{s,t}\}$ has a Lipschitz continuous extension to $T \times T$, so the preceding corollary implies that $Y_{s,t}$ extends continuously to all of $T \times T$.

2.2. Maximum of a Gaussian process.

Theorem 1 implies that if the metric entropy criterion (2.1) holds then the Gaussian process has a version with continuous sample paths, and for this version the maximum must be attained with probability one. A minor variation of the argument used to prove Theorem 1 shows that the *expectation* of the max is bounded by a multiple of the metric entropy integral.

Theorem 9. *There is a universal constant $K < \infty$ such that for any non-degenerate, centered Gaussian process $\{X_t\}_{t\in T}$ and any countable dense subset $D \subseteq T$,*

$$E \sup_{t\in D} X_t \leq K \int_0^{\operatorname{diam}(T)/2} \sqrt{\log N(\varepsilon)}\; d\varepsilon \tag{6}$$

This is stated so that it holds for any version, whether or not it is continuous. However, it is easy to see, given Theorem 1, that the statement is equivalent to the statement that for a *continuous* version,

$$E \max_{t \in T} X_t \leq K \int_0^{\operatorname{diam}(T)/2} \sqrt{\log N(\varepsilon)}\, d\varepsilon \tag{7}$$

The high concentration of sup around its mean is an easy consequence of the Gaussian concentration inequality.

Theorem 10. *Let $\{X_t\}_{t \in T}$ be a centered Gaussian process on a countable set T that is almost surely bounded. Write $X^* = \sup_{t \in T} X_t$. If $\sigma_t^2 := \sup_{t \in T} E X_t^2 < \infty$, then $E X^* < \infty$ and for any $u > 0$*

$$P\left(X^* \geq E X^* + u\right) \leq \exp\left(-u^2 / 2\sigma_t^2\right) \tag{8}$$

Proof. Consider first the case where $T = [m]$ is finite, and without loss of generality assume that $\sigma_t^2 \leq 1$. Then $X := \{X_t\}_{t \in T}$ is just a mean-zero Gaussian random vector, and X^* is the value of the maximum coordinate. This has finite expectation, because, for instance it is always bounded by the absolute value. Furthermore, if $k \leq m$ is the rank of the covariance matrix then there is a linear transformation $A \colon \mathbb{R}^k \to \mathbb{R}^m$ and a standard Gaussian random vector Y in $\mathbb{R}^k$ such that $X = AY$. Because $\sigma_t^2 \leq 1$, the components X_j of X have variances bounded by 1, and so the rows of A all have length bounded by 1. Consequently, the mapping $A^* \colon \mathbb{R}^k \to \mathbb{R}$ that takes $y \in \mathbb{R}^k$ to the largest coordinate of the m-vector Ay is 1–Lipschitz. Hence, the inequality (2.5) is an immediate consequence of the Gaussian concentration theorem.

Now let T be countable, and let $T_1 \subseteq T_2 \subseteq \ldots$ be finite subsets of T whose union is T. Set $\mu_n = E \max_{t \in T_n} X_t$; since the sets T_n are nested, the sequence μ_n is nondecreasing. There are two possibilities: either $\lim \mu_n < \infty$ or $\lim \mu_n = \infty$. But if $\lim \mu_n = \infty$ then it must be the case that $X^* = \infty$ almost surely, because $X^* \geq \max_{t \in T_n} X_t$, and since (2.5) holds for each T_n, the maximum of X on T_n is highly concentrated about μ_n. (Exercise: Fill in the details.) Thus, if $X^* < \infty$ almost surely then $\lim \mu_n < \infty$, and by monotone convergence

$$E X^* = E \lim_{n \to \infty} \max_{t \in T_n} X_t = \lim_{n \to \infty} \mu_n < \infty \tag{9}$$

Finally, to prove the concentration inequality, observe that for any $u > 0$

$$\begin{aligned} P\left(X^* \geq E X^* + u\right) &= \lim_{n \to \infty} P\left(\max_{t \in T_n} X_t \geq E X^* + u\right) \\ &\leq \lim_{n \to \infty} P\left(\max_{t \in T_n} X_t \geq \mu_n + u\right) \\ &\leq e^{-\frac{u^2}{2\sigma_t^2}} \end{aligned} \tag{10}$$ □

3. Reproducing Kernel Hilbert Space of a Gaussian Process

Let $R(s,t)$ be the covariance function of a centered, non-degenerate Gaussian process $\{X_t\}_{t\in T}$, and let L_2^X be the closure, relative to the L_2–norm, of the vector space consisting of all finite linear combinations of the random variables X_t. Since L_2–limits of centered Gaussian random variables are centered Gaussian random variables, every random variable in L_2^X is Gaussian. The *reproducing kernel Hilbert space* (abbreviated RKHS) associated to the Gaussian process $\{X_t\}_{t\in T}$ is a Hilbert space of real-valued functions on T that is naturally isometric to L_2^X. The isometry between these Hilbert spaces leads to useful spectral representations of the Gaussian process, notably the Karhunen-Loeve representation.

3.2. Reproducing Kernel Hilbert Space.

Definition 11. *A symmetric, real-valued function $R\colon T\times T\to\mathbb{R}$ is said to be a positive definite kernel if for every finite subset $F\subseteq T$ the $|F|\times|F|$ matrix $(R(s,t))_{s,t\in F}$ is positive definite.*

Definition 12. *The covariance function of any nondegenerate Gaussian process has this property. Every positive definite kernel R on T induces a canonical metric d on T by*

$$d(s,t)=\sqrt{R(t,t)+R(s,s)-2\,R(s,t)} \tag{11}$$

Lemma 13. *The definition (3.5) specifies a metric on T.*

Proof. This is most easily proved by making a connection with Gaussian processes. To prove that d is a metric on T it suffices to prove that it is a metric on any finite subset of T (because the defining properties of a metric – symmetry, positivity, and the triangle inequality – involve only two or three points at a time). Now if F is a finite subset of T then the matrix $(R(i,j))_{i,j\in F}$ is a symmetric, positive definite $|F|\times|F|$ matrix, and hence is the covariance matrix of a centered, nondegenerate Gaussian random vector $(X_i)_{i\in F}$. By construction,

$$d(i,j)^2=E\,|X_i-X_j|^2$$

so d is the canonical metric of the Gaussian process $(X_i)_{i\in F}$. But the canonical metric of a non-degenerate Gaussian process is a genuine metric.

Assume for the remainder of this section that $R(s,t)$ is a positive definite kernel on T. For each $s\in T$ let $R_s\colon T\to\mathbb{R}$ be the s–section of R, that is, the function defined by $R_s(t)=R(s,t)$. Let H_0 be the vector space of real-valued functions on T consisting of all finite linear combinations $\sum a_i\,R_{s_i}$. Define an inner product on H_0 as follows:

$$\Big\langle\sum a_i\,R_{s_i},\sum b_j\,R_{t_j}\Big\rangle=\sum a_i\,b_j\,R(s_i,t_j)$$

This is clearly bilinear, and since R is assumed to be positive definite it defines an inner product (as you should check!). This inner product is designed so that the mapping $s \to R_s$ is an *isometry* of T (with the canonical metric) into H_0 (with the metric induced by the inner product (3.6)): in particular, for any $s, t \in T$

$$\|R_s - R_t\|^2 = R(t,t) + R(s,s) - 2\,R(s,t) = d(s,t)^2 \tag{12}$$

Observe that when R is the covariance function of a centered Gaussian process $\{X_t\}_{t\in T}$, the inner product (3.6) makes the mapping $T\colon H_0 \to L^2(P)$ defined by

$$T\big(\sum a_i\, R_{s_i}\big) = \sum a_i\, X_{s_i} \tag{13}$$

a linear isometry.

The pairing of the inner product (3.6) with the positive definite kernel R is explicitly made so that the kernel $R(s,t)$ will a *reproducing kernel* relative to the inner product, in the following sense: for any function $\varphi \in H_0$ and any $t \in T$,

$$\varphi(t) = \langle R_t, \varphi \rangle \tag{14}$$ □

Definition 14. *The reproducing kernel Hilbert space H associated with the covariance kernel R is the closure of H_0 with respect to the norm induced by the inner product (3.6).*

In the special case where R is the covariance function of a centered Gaussian process $\{X_t\}_{t\in T}$, Proposition 3.1 implies that the linear isometry T defined by (3.8) extends in a unique way to all of H, since by construction H_0 is a dense linear subspace of H.

Lemma 15. *If T is separable relative to the canonical metric then so is the reproducing kernel Hilbert space H, and if T is compact then set $\{R_s\}_{s\in T}$ is compact, and therefore bounded, in H.*

Proof. By construction, the mapping $s \to R_s$ is an isometry. Therefore, if T is compact (separable) then the set $\{R_s\}_{s\in T}$ is compact (separable) in H. If the set $\{R_s\}_{s\in T}$ is separable, it has a countable dense subset D, and if this is the case then the (countable) set of rational linear combinations of elements of D is dense in the reproducing kernel Hilbert space H. □

Lemma 16. *The set $\{R_s\}_{s\in T}$ is bounded in the reproducing kernel Hilbert space H if and only if $\sup_{t\in T} R(t,t) < \infty$. If this is the case then*

(a) the function $(s,t) \to R(s,t)$ is Lipschitz in each variable s, t; and

(b) for every element $\varphi \in H$ the function $t \to \langle R_t, \varphi \rangle$ is a Lipschitz continuous function on T, with Lipschitz constant bounded by $\|\varphi\|$ and sup-norm bounded by $\|\varphi\| \sup_{t\in T} \|R_t\|$.

Thus, the elements $\varphi \in H$ can be identified with Lipschitz functions $t \to \langle R_t, \varphi \rangle$ on T.

Proof. By the reproducing property (or equivalently the definition of the inner product (3.6)),

$$\|R_s - R_t\|^2 = \langle R_s - R_t, R_s - R_t\rangle = R(t,t) + R(s,s) - 2\,R(s,t) = d(s,t)^2 \tag{15}$$

Assume now that $C = \sup_{t\in T} \|R_t\| < \infty$. By the reproducing property and the Cauchy-Schwartz inequality, for any $s_1, s_2 \in T$,

$$|R(s_1,t) - R(s_2,t)| = |\langle R_t - R_{s_2}, R_t\rangle| \le |\langle R_s - R_{s_2}, R_t\rangle| \le d(s_1,s_2)\|R_t\|$$

$$= C\,d(s_1,s_2) \tag{16}$$

consequently, $R(s,t)$ is Lipschitz in s, with Lipschitz constant bounded by C. Since $R(s,t) = R(t,s)$, it follows that R is Lipschitz in the second variable t as well.

Finally, let φ be any element of the reproducing kernel Hilbert space H, and consider the mapping $t \to \langle \varphi, R_t\rangle$. For any two points $s, t \in T$,

$$|\langle R_t, \varphi\rangle - \langle R_s, \varphi\rangle| = |\langle R_t - R_s, \varphi\rangle| \le d(s,t)\|\varphi\| \tag{17}$$

This implies that the function $t \to \langle R_t, \varphi\rangle$ is Lipschitz, with Lipschitz constant bounded by $\|\varphi\|$. Furthermore, by Cauchy-Schwartz, for any $t \in T$,

$$|\langle R_t, \varphi\rangle| \le \|\varphi\| \sup_{t\in T} \|R_t\| \tag{18}$$ □

Note 17. Lest you be confused about the identification of elements $\varphi \in H$ with functions $t \to \langle R_t, \varphi\rangle$: keep in mind that H was constructed as the abstract completion of the inner product space H_0, so (technically) its elements are equivalence classes of Cauchy sequences in H_0, not functions on T (even though the elements of H_0 are functions on T). Nevertheless, when the hypotheses of Lemma 3.5 hold, we will act as if the elements $\varphi \in H$ are Lipschitz functions and write $\varphi(t) = \langle R_t, \varphi\rangle$.

Proposition 18. *If T is compact relative to the canonical metric of the Gaussian process then for any (countable) orthonormal basis $\{\psi_n\}_{n\ge 1}$ and any $s, t \in T$*

$$R(s,t) = \sum_{n\ge 1} \langle R_s, \psi_n\rangle\, \psi_n(t) = \sum_{n\ge 1} \psi_n(s)\, \psi_n(t) \tag{19}$$

and the convergence is uniform for $s, t \in T$.

Note 19. This is a variant of Mercer's theorem. The hypothesis that T is compact ensures that the reproducing kernel Hilbert space is separable, and hence that every orthonormal basis is countable (or has cardinality $|T|$ is T is finite). If T is finite the expansion (3.10) follows from the spectral theorem for finite symmetric matrices, so we will restrict attention to the case where T is compact but infinite.

Proof. By construction, the mapping $s \to R_s$ is an isometry. Therefore, if T is compact then by Lemma 3.5, R is Lipschitz in each variable. This implies that the mapping $s \to R(s, s)$ is Lipschitz, and that $\sup_{s \in T} R(s, s) < \infty$.

If T is compact and infinite then every orthonormal basis for the reproducing kernel Hilbert space is countable. Let $\{\psi_n\}_{n \geq 1}$ be any orthonormal basis. By Parseval's theorem, for any $s \in T$ the s–section R_s has the expansion

$$R_s = \sum_{n \geq 1} \langle R_s, \psi_n \rangle \, \psi_n = \sum_{n \geq 1} \psi_n(s) \, \psi_n \tag{20}$$

where the convergence is in the H–norm. (The second identity follows from the reproducing property of the kernel.) Moreover, for any $s \in T$,

$$R(s, s) = \|R_s\|^2 = \sum_{n \geq 1} \langle R_s, \psi_n \rangle^2 = \sum_{n \geq 1} \psi_n(s)^2 \tag{21}$$

Since T is compact, the function $s \to R(s, s)$ is Lipschitz, by Lemma 3.5, and therefore continuous. Moreover, the functions $\psi_n(s)^2$ are (Lipschitz) continuous, also by Lemma 3.5, and obviously nonnegative. Consequently, Dini's theorem[1] implies that the series $\sum_{n \geq 1} \psi_n(s)^2$ converges to $R(s, s)$ uniformly for $s \in T$.

To complete the proof we will show that the uniform convergence of the series $\sum_{n \geq 1} \psi_n(s)^2$ implies that the series $\sum_{n \geq 1} \psi_n(s) \, \psi_n(t)$ converges uniformly for $(s, t) \in T \times T$. The uniform convergence follows by Cauchy-Schwartz, because for any $m \geq 1$,

$$\left| \sum_{n \geq m} \psi_n(s) \, \psi_n(t) \right| \leq \sqrt{\sum_{n \geq m} \psi_n(s)^2 \sum_{n \geq m} \psi_n(t)^2} \tag{22}$$

since $\sum_{n \geq 1} \psi_n(s)^2$ converges uniformly for $s \in T$, the right side of the inequality can be made uniformly small by choosing m large. To see that the limit of the series $\sum_{n \geq 1} \psi_n(s) \, \psi_n(t)$ is $R(s, t)$, we use the reproducing property of the kernel together with the H–convergence (3.11). This implies that

$$R(s, t) = \langle R_s, R_t \rangle = \sum_{n \geq 1} \psi_n(s) \langle \psi_n, R_t \rangle = \sum_{n \geq 1} \psi_n(s) \, \psi_n(t) \tag{23}$$ □

1. Dini's theorem states that if $f_n(t)$ is a nondecreasing sequence of nonnegative continuous functions on a compact space T such that $f_n(t) \uparrow f(t)$ for every $t \in T$, where f is a continuous function, then the convergence is uniform.

Proposition 20. *If T is compact relative to the canonical metric then on any probability space that supports a sequence $\{\xi_n\}_{n\geq 1}$ of independent $N(0,1)$ random variables there exists a centered Gaussian process $\{X_t\}_{t\in T}$ with covariance function R. For any orthonormal basis $\{\psi_n\}_{n\geq 1}$ and any $t\in T$ the random variable X_t is the almost sure limit of the series*

$$X_t = \sum_{n\geq 1} \xi_n \psi_n(t) \tag{24}$$

Proof. If T is separable then so is the reproducing kernel Hilbert space H, by the same argument as in the proof of Proposition 3.6. Consequently, H has a countable orthonormal basis ψ_n, and for each $s\in T$ the s–section R_s has the expansion (3.11), and has norm $\|R_s\|$ given by (3.12). This implies that the series $\sum_{n\geq 1}\psi_n(s)^2$ converges to a finite limit.

Suppose that $\{\xi_n\}_{n\geq 1}$ is a sequence of independent $N(0,1)$ random variables. Then for any $s\in T$ the series

$$X_s := \sum_{n\geq 1} \psi_n(s)\,\xi_n \tag{25}$$

converges in $L^2(P)$ and almost surely, since the variances $\psi_n(s)^2$ are summable. (The sequence of partial sums is an L^2–bounded martingale, and a basic result of martingale theory asserts that any such martingale converges both a.s. and in $L^2(P)$ by the usual arguments, the limit random variable X_s is centered Gaussian, and the process $\{X_s\}_{s\in T}$ is Gaussian. The covariance function is

$$E\,X_s\,X_t = E\sum_n\sum_m \psi_n(s)\,\psi_m(t)\,\xi_n\,\xi_m = \sum_{n\geq 1}\psi_n(s)\,\psi_n(t) = R\,(s,t) \tag{26}$$ □

Theorem 21. *(Itô-Nisio) Let $\{X_t\}_{t\in T}$ be a centered Gaussian process with covariance function R and continuous sample paths. Assume also that T is compact relative to the canonical metric d. Let $\{\psi_n\}_{n\geq 1}$ be any orthonormal basis for the reproducing kernel Hilbert space, and denote by $\xi_n = T\psi_n$ the image of ψ_n under the canonical isometry T (see equation (3.8)). Then the random variables ξ_n are independent Normal(0,1), and*

$$X_t = \sum_{n\geq 1} \xi_n \psi_n(t) \tag{27}$$

With probability one, the series converges uniformly in t.

The main point of the theorem is the uniform convergence. The hypothesis that the Gaussian process has a version with continuous paths is crucial, as there are Gaussian processes with the property that every version is discontinuous at every point with probability one (see Example 2.1 above). Theorem 4 is somewhat harder to prove than the other results of this section, so I will omit it and refer you to the original paper of Itô and Nisio.

Note 22. Since the random variables ξ_n are i.i.d. the series $\sum \xi_n^2 = \infty$ a.s. Therefore, the expansion (3.14) does not converge in H, and in particular the sample paths $t \to X_t$ of the Gaussian process are a.s. not elements of H. (Keep this in mind when reading Example 3.2 below.)

3.3. Examples of RKHS.

Example 23. Recall that the covariance function of the standard Wiener process $\{W_t\}_{t\in[0,1]}$ is $R(s,t) = s \wedge t$, where $\wedge$ denotes min. Each s–section R_s has the property that $R_s(0) = 0$, so every function in H_0 takes the value 0 at $t = 0$. Each s–section R_s has derivative R_s' equal to the indicator function $I_{[0,s]}$ of the interval $[0,s]$, so every $\varphi \in H_0$ is differentiable at all but finitely many points, and the derivative is bounded. Hence, for every $\varphi \in H_0$,

$$\langle R_s, \varphi \rangle = \varphi(s) = \int_0^1 R_s'(t)\, \varphi'(t)\, dt \tag{28}$$

It follows that the inner product (3.6) on H_0 coincides with the usual inner product of derivatives, that is,

$$\langle \varphi, \psi \rangle = \int_0^1 \varphi'(s)\, \psi'(s)\, ds \tag{29}$$

Since linear combinations of intervals $I_{[0,s]}$ are dense in $L^2[0,1]$, it follows that H_0 is dense in the space of all functions $\varphi\colon [0,1] \to \mathbb{R}$ with L^2–derivative φ' and $\varphi(0) = 0$. Consequently, this space is the reproducing kernel Hilbert space for Brownian motion. It is sometimes called the Cameron-Martin Hilbert space.

4. Mutual Singularity and Absolute Continuity

1 Distinguishibility Of Measures

Let $\{X_t\}_{t\in T}$ be a centered, nondegenerate Gaussian process with (positive definite) covariance function $R(s,t)$, and assume that the index set T is compact relative to the canonical metric d. Assume also that the sample paths $t \to X_t$ are continuous. Given a non-random, continuous function $f\colon T \to \mathbb{R}$ – the "signal" – when can one distinguish between $\{X_t + f_t\}_{t\in T}$ and $\{X_t\}_{t\in T}$?

More precisely, let P and Q be probability measures on the measurable space $(\omega, \mathcal{F})$ such that under P the process $\{X_t\}_{t\in T}$ is centered Gaussian with covariance function R, while under Q the process $\{X_t - f_t\}_{t\in T}$ is centered Gaussian with covariance function R. Assume that the σ–algebra $\mathcal{F}$ is generated by the random variables $\{X_t\}_{t\in T}$. Recall that there is a linear isometry $J_X\colon H \to L^2(P)$ that maps R_t to the coordinate variable X_t (see equation (3.7)).

Theorem 24. *The probability measures P, Q are mutually singular on $(\Omega, \mathcal{F})$ unless f is an element of the reproducing kernel Hilbert space H of R. If $f \in H$ then P, Q are mutually absolutely continuous, with likelihood ratio (Radon-Nikodym derivative)*

$$\frac{dQ}{dP}|_{\mathcal{F}} = \exp\{J_X(f) - \|f\|^2/2\} \tag{30}$$

Remark 25. In the special case where $\{X_t\}_{t\in T}$ is the Wiener process on $T = [0,1]$ this theorem was discovered and proved by Cameron and Martin around 1945. In this case the isometry J_X is the usual Wiener integral. Later Girsanov extended the Cameron-Martin formula (4.2) for the likelihood ratio to certain *random* functions f given as stochastic integrals.

Before turning to the proof of Theorem 5 let's look at the related but at least superficially simpler problem of distinguishing signal plus noise from noise for *white noise*. Here $T = \mathbb{N}$, and the covariance function is the simplest of all possible covariance functions, to wit, $R(i, j) = \delta_{i,j}$; thus, the Gaussian process $\{X_n\}$ consists of independent, identically distributed standard Gaussian random variables. The "signal" is a fixed non-random sequence $\{a_n\}_{n\in\mathbb{N}}$ of real numbers. Let the probability measures μ and ν be the joint distributions of the sequences $\{X_n\}$ and $\{X_n + a_n\}$, respectively.

Theorem 26. *(Kakutani) The measures μ and ν are absolutely continuous if*

$$\sum_{n\geq 1} a_n^2 < \infty \tag{31}$$

, and mutually singular if

$$\sum_{n\geq 1} a_n^2 = \infty \tag{32}$$

If $\sum_{n\geq 1} a_n^2 < \infty$ then the series $\sum_{n\geq 1} a_n y_n$ converges almost surely under μ and

$$\frac{d\nu}{d\mu}|_{(y_n)_{n\geq 1}} = L_\infty = e^{\left(\sum_{n\geq 1} a_n y_n - \frac{1}{2}\sum_{n\geq 1} a_n^2\right)} \tag{33}$$

Proof. First some generalities. Let μ, ν be two probability measures on $(\mathbb{R}^\infty, \mathcal{F}_\infty)$, where $\mathcal{F}_\infty$ is the usual Borel field, that is, the smallest σ–algebra that contains all cylinder sets. For each $n \in \mathbb{N}$ denote by $\mathcal{F}_n$ the σ–algebra generated by the first n coordinate variables. Suppose that the restrictions of μ and ν to $\mathcal{F}_n$ are mutually absolutely continuous for each n (this is the case, for instance, when μ and ν are the laws of any two nondegenerate Gaussian processes, or when μ and ν are product measures where each factor is absolutely continuous relative to Lebesgue measure on $\mathbb{R}$). Then the sequence

$$L_n := \left(\frac{d\nu}{d\mu}\middle| \mathcal{F}_n\right) \tag{34}$$

of likelihood ratios is a positive martingale under μ relative to the filtration $\{\mathcal{F}_n\}_{n\geq 1}$, and hence converges to a nonnegative limit L_∞ almost surely. (Exercise: Verify that L_n is a martingale.) There are two interesting cases:

a) $E_\mu L_\infty = 1$

b) $\mu\{L_\infty = 0\} = 1$

These are not the only possibilities, in general, but in both the Gaussian cases and in the case of product measures either (a) or (b) will hold. In case (a) the martingale L_n is closed, that is, $L_n = E(L_\infty|\mathcal{F}_n)$ for each n and $L_n \to L_\infty$ in L^1. In this case the measures μ and ν are absolutely continuous on $\mathcal{F}_\infty$, with Radon-Nikodym derivative L_∞, because for any $F \in \mathcal{F}_n$,

$$E_\mu L_\infty 1_F = E_\mu L_n 1_F = \nu(F) \tag{35}$$

and so the measures ν and $F \mapsto E_\mu L_\infty 1_F$ agree on $\bigcup_{n\geq 1}\mathcal{F}_n$. In case (b), on the other hand, the measures μ and ν are singular on $\mathcal{F}_\infty$. To see this, first note that if $L_n \to 0$ a.s. under μ then there is a sequence of constants $\varepsilon_n > 0$ such that $\varepsilon_n \to 0$ and such that with μ–probability 1,

$$\mu\left(L_n \leq \varepsilon_n \text{eventually}\right) = 1 \tag{36}$$

(*Exercise:* Prove this.) Define τ_m to be the first time $n \geq m$ that $L_n \leq \varepsilon_n$, and let F_m be the event $\{\tau_m < \infty\}$. These events are nested: $F_m \supseteq F_{m+1}$ for each m. By construction, $\mu(F_m) = 1$ for every $m \geq 1$, and hence $\mu(\bigcap_m F_m) = 1$. But for each $m \geq 1$,

$$\nu(F_m) = \sum_{n=m}^{\infty} \nu\left(\tau_m = n\right) = \sum_{n=m}^{\infty} E_\mu L_n 1_{\{\tau_m = n\}} = E_\mu L_{\tau_m} \leq \varepsilon_m \tag{37}$$

Consequently, $\nu(\bigcap_m F_m) = 0$.

Now to Kakutani's theorem: Under μ the coordinate variables y_n are i.i.d. standard Gaussian, whereas under ν they are independent Gaussians with means a_n and variances 1. Thus, the likelihood ratio on $\mathcal{F}_n$ is

$$L_n = e^{A_n z_n - \frac{A_n^2}{2}} \quad \text{where} \quad A_n^2 = \sum_{j=1}^{n} a_j^2 \quad \text{and} \quad z_n = \sum_{j=1}^{n} \frac{a_j y_j}{A_n} \tag{38}$$

If $A = \sum a_j^2 < \infty$ then the series $\sum_{j=1} a_j y_j$ converges almost surely and in $L^2(\mu)$, and the limit random variable is centered Gaussian with variance A. Consequently, $L_n \to L_\infty$ where half of the theorem. Suppose then that $\sum a_j^2 = \infty$. Under ν the random variable z_n is standard Gaussian, and so

$$\lim_{n\to\infty} \mu\left\{A_n z_n > \frac{A_n^2}{4}\right\} = 0 \tag{39}$$

This implies that $L_n \to 0$ in μ–probability. But the martingale convergence theorem implies that $L_n \to L_\infty$ almost surely under μ, and so it follows that $L_\infty = 0$ a.s. μ. □

Proof of Theorem 5. Since T is compact it has a countable dense subset $D = \{s_1, s_2, \ldots\}$. The strategy of the proof will be to reduce the problem to the situation covered by Kakutani's theorem by applying Gram-Schmidt to the random variables $X_{s_1}, X_{s_2}, \ldots$.

Assume then that under P the process $\{X_t\}_{t\in T}$ is centered Gaussian with covariance function $R(s,t)$, and that under Q the process $\{X_t - f_t\}_{t\in T}$ is centered Gaussian with covariance function $R(s,t)$. Assume also that under both P and Q the sample paths $t \to X_t$ are continuous; this implies that the σ–algebra $\mathcal{F}$ generated by the entire family $\{X_t\}_{t\in T}$ coincides with the σ–algebra $\mathcal{F}$ generated by $\{X_t\}_{t\in D}$, and so

$$\mathcal{F} = \sigma\left(\bigcup_{n\geq 1} \mathcal{F}_n\right) \tag{40}$$

where $\mathcal{F}_n = \sigma((X_s)_{s\in D_n})$. Let $\xi_1, \xi_2, \ldots$ be the random variables obtained by applying the Gram-Schmidt algorithm to the sequence $X_{s_1}, X_{s_2}, \ldots$, that is,

$$\xi_1 = \frac{X_{s_1}}{\sqrt{R(s_1, s_1)}} \tag{41}$$

$$\xi_2 = \frac{X_{s_2} - \frac{R(s_2,\, s_1)\, X_{s_1}}{R(s_1,\, s_1)}}{\sqrt{\left(R(s_1, s_1) - \frac{R(s_2,\, s_1)^2}{R(s_1,\, s_1)}\right)}} \tag{42}$$

etc.

Since we have assumed that the covariance kernel $R(s,t)$ is positive definite, the random variables $\{X_{s_j}\}_{j\geq 1}$ are linearly independent in $L^2(P)$, and hence the Gram-Schmidt mapping (4.4) is well-defined (that is, there are no divisions by 0). Observe that for each $n \geq 1$ the Gram-Schmidt equations (4.4) can be written in matrix form:

$$\begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix} = T_n \begin{pmatrix} X_{s_1} \\ X_{s_2} \\ \vdots \\ X_{s_n} \end{pmatrix} \tag{43}$$

where for each n the matrix T_n is lower triangular and invertible. Consequently, for each n,

$$\mathcal{F}_n = \sigma(\{\xi_i\}_{i\leq n})$$

Thus, the random variables $\{\xi_n\}$ generate the σ-algebra $\mathcal{F}$. Note also that the entries of the matrices T_n are determined solely by the entries $R(s_i, s_j)$ with $i, j \le n$, and that for each n the $n \times n$ minor of T_{n+1} is T_n, that is,

$$T_n = \begin{pmatrix} b_{1,1} & 0 & 0 & \cdots & 0 \\ b_{2,1} & b_{2,2} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ b_{n,1} & b_{n,2} & b_{n,3} & \cdots & b_{n,n} \end{pmatrix} \tag{44}$$

for fixed scalars $b_{i,j}$.

By construction, the random variables $\{\xi_n\}_{n\ge1}$ are independent, identically distributed standard Gaussians under P. Under Q, however, they are independent Gaussians with common variance 1, but with means $a_n = E_Q \xi_n$ depending on the function f. These means can be computed by taking E_Q–expectations on both sides in (4.5): this implies that for each n,

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = T_n \begin{pmatrix} f_{s_1} \\ f_{s_2} \\ \vdots \\ f_{s_n} \end{pmatrix} \tag{45}$$

In particular, for each n the vector $(a_i)_{i\le n}$ is gotten by applying the Gram-Schmidt transformation to the vector $(f_{s_i})_{i\le n}$.

Suppose that $f \in H$. Let $J_X : H \to L^2(P)$ be the canonical linear isometry from the reproducing kernel Hilbert space H into $L^2(P)$ (recall that this is determined by the identification $J_X(R_t) = X_t$), and let $\psi_n = J_X^* \xi_n$. Since J_X is a linear isometry, the sequence $\{\psi_n\}_{n\ge1}$ constitutes an orthonormal basis of H, and the functions ψ_n are obtained from the sequence $R_{s_1}, R_{s_2}, \ldots$ by Gram-Schmidt, and for each n the linear transformation taking $(R_{s_j})_{j\le n}$ to $(\psi_j)_{j\le n}$ is the same as that taking $(X_{s_j})_{j\le n}$ to $(\xi_j)_{j\le n}$. This implies that

$$a_n = \langle f, \psi_n \rangle \tag{46}$$

Since $f = \sum_{n\ge1} \langle f, \psi_n \rangle \psi_n$, it follows that $\sum_{n\ge1} a_n^2 = \|f\|^2 < \infty$. Thus, Kakutani's theorem implies that in this case the measures P and Q are absolutely continuous, and the likelihood ratio is

$$\frac{dQ}{dP}|_{\mathcal{F}} = e^{\left(\sum_{n\ge1} a_n \xi_n - \frac{\sum_{n\ge1} a_n^2}{2}\right)} \tag{47}$$

The formula (4.2) follows, because

$$J_X(f) = \sum_{n\ge1} a_n \xi_n \quad \text{and} \quad \|f\|^2 = \sum_{n\ge1} a_n^2 \tag{48}$$

Now suppose that $f \notin H$. For each $n \geq 1$, let $f^n \in H$ be the unique linear combination in $\text{span}\{R_{s_j}\}_{j \leq n}$ such that $f^n(s_i) = f(s_i)$ for each $i \leq n$. There is such a linear combination, because for each n the matrix $R(s_i, s_j)_{i,j \leq n}$ is positive definite, and hence invertible. For each n the means $(a_i)_{i \leq n}$ are related to the entries $(f_{s_i})_{i \leq n}$ by (4.6), and so

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = T_n \begin{pmatrix} f_{s_1} \\ f_{s_2} \\ \vdots \\ f_{s_n} \end{pmatrix} \tag{49}$$

But since f^n is a linear combination of the functions $(R_{s_j})_{j \leq n}$, it follows that for $1 \leq m \leq n$,

$$a_m = \langle f^n, \psi_m \rangle \tag{50}$$

and so

$$\sum_{j=1}^{n} a_j^2 = \| f^n \|^2 \tag{51}$$

Now if it were the case that $\sum_{j=1}^{\infty} a_j^2 = \| f^n \|^2 < \infty$ then $g = \sum_{n \geq 1} a_n \psi_n \in H$, and the last two equations would imply that for each n the function f^n is the orthogonal projection of g on the subspace spanned by the functions $(R_{s_j})_{j \leq n}$. But this would imply that $g(s_i) = f(s_i)$ for every $i \geq 1$, and since the sequence $(s_i)_{i \geq 1}$ is dense in T it would follow that $f = g$ everywhere. Since we have assumed that $f \notin H$ this cannot be the case. Therefore, $\sum_{n=1}^{\infty} a_n^2 = \infty$, and Kakutani's theorem implies that P and Q are mutually singular.

2 Intuition

Mutual singularity between measures, like P and Q in the Gaussian process context, means that there exist disjoint sets in the underlying space, each fully supported by one measure and null for the other. Basically, the measures live on completely separate parts of the probability space, making them incompatible or "orthogonal."

Absolute continuity, on the other hand, is when one measure (like Q) is completely contained within another (like P). If f is in the reproducing kernel Hilbert space (H) associated with the Gaussian process, then P and Q are absolutely continuous. This relationship is characterized by the existence of a Radon-Nikodym derivative $d\,Q\,/\,d\,P$, giving the precise "scaling" of probabilities from P to Q across events.

The intuition here revolves around how the modifications (f_t) to the Gaussian process X_t under measure Q compared to P affect the likelihood of observing certain outcomes. If f belongs to H, it aligns well enough with the Gaussian structure of X_t under P that Q just reweights these probabilities (absolute continuity). If f does not belong to H, the modifications it introduces are too orthogonal (in the Hilbert space sense) to the original process, making P and Q live on different "worlds" (mutual singularity).

University of Chicago, Department of Statistics, USA