1 Introduction

Gaussian distributions are important building blocks in statistical modeling. As modeling tools, they arise naturally from first principles reasonings based off the central limit theorem. They also provide the most natural way to incorporate dependency between multiple quantities. When there are infinitely many such quantities that can be viewed as a map or a function from an input space to an output space, the notion of Gaussianity continues to provide useful tools for constructing joint probability models. In particular, there is a notion of a Gaussian probability distribution on infinite dimensional spaces such as $\mathbb{R}^S$, the space of all functions $f : S \to \mathbb{R}$, where $S$ is a subset of an Euclidean space. Below we build up a general framework of the notion of Gaussianity on arbitrary spaces.

2 Gaussian random elements

As a starting point we only assume the concept of a univariate normality. A random variable $X \in \mathbb{R}$ is said to be normally distributed if there exist $\mu \in \mathbb{R}$, $\sigma > 0$ such that the pdf of $X$ is $f(x) = \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}/\sqrt{2\pi\sigma^2}$, $x \in \mathbb{R}$. The distribution of $X$ is $N(\mu, \sigma^2)$. For completion, we allow the notation $N(0, 0)$ to mean the degenerate distribution under which $X = 0$ with probability 1.

2.1 Formal definition

Below we define the concept of a Gaussian random element in a Banach space $\mathcal{F}$, a (complete) normed linear space with a norm $\| \cdot \|$. Two specific examples of interest to us are

- $\mathcal{F} = \mathbb{R}^p$ for some $p \geq 1$ equipped with the Euclidean norm
- $\mathcal{F} = C(S)$, the space of all continuous functions from a compact $S \subset \mathbb{R}^d$ to $\mathbb{R}$, equipped with the supremum norm $\|f\|_\infty = \sup_{s \in S} |f(s)|$.

Notice that the first space is also a Hilbert space while the second is not. Additionally, we would occasionally look at

- $\mathcal{F} = L_2(S, \nu)$ for some $S \subset \mathbb{R}^d$ and a measure $\nu$ on $S$, equipped with the $L_2$ norm $\|f\|_2 = \left\{\int_S f(s)^2\nu(ds)\right\}^{1/2}$.

For our discussion this space will mostly serve as a caveat against blind generalizations.

Recall that linear operators on a Banach space $\mathcal{F}$ are linear maps $L : \mathcal{F} \to \mathbb{R}$. A linear operator is continuous if and only if it is bounded, i.e., the operator norm
\[ \|L\| := \sup_{f \in \mathcal{F}} |Lf|/\|f\| \] is finite. The set of all continuous/bounded linear operators is called the dual space \( \mathcal{F}^* \) of \( \mathcal{F} \). The dual space itself is a Banach space under the operator norm.

**Definition 1.** A random element \( W \) in a Banach space \( (\mathcal{F}, \| \cdot \|) \) is said to have a Gaussian law if for every \( L \in \mathcal{F}^* \), the scalar random variable \( LW \) is normally distributed. \( W \) is called centered if \( LW \) has mean zero for every such \( L \). \( W \) is said to be non-degenerate if \( LW = 0 \) with probability one only when \( L = 0 \).

When \( \mathcal{F} = \mathbb{R}^p \), this defines the \( p \)-variate Gaussian distribution. In this case, any continuous linear map \( L \) is an inner product operation with respect to some vector \( a \in \mathbb{R}^p \), \( L : f \mapsto a^Tf \). The law of \( W \) is uniquely characterized by \( \mu = \mathbb{E}(W) \) and \( \Sigma = \mathbb{V}(W) \). Clearly, \( \Sigma \) is a non-negative definite matrix, i.e., for any \( u \in \mathbb{R}^p \), \( u^T \Sigma u \geq 0 \).

When \( \mathcal{F} \subset \mathbb{R}^S \), any random element \( W \) in \( \mathcal{F} \) is both a random function \( W : S \to \mathbb{R} \) and a stochastic process \( W = (W(s) : s \in S) \). If \( W \) is Gaussian in \( \mathcal{F} \), we can as well call it a Gaussian process.

When \( \mathcal{F} = C(S) \) for some compact \( S \subset \mathbb{R}^d \), for every \( s \in S \), the evaluation function \( \delta_s : f \mapsto f(s) \) is a continuous linear operator (of norm 1). Therefore, \( W(s) = \delta_s W \) is normally distributed for any \( s \in S \). Moreover, for any \( s_1, \ldots, s_n \in S \), \( X = (W(s_1), \ldots, W(s_n))^T \) satisfies the Gaussianity condition for random elements in \( \mathbb{R}^n \), and hence \( X \) is a multivariate normal random variate.

We will show that the law of any centered Gaussian element \( \hat{W} \in C(S) \) is uniquely characterized by the covariance function \( k : S \times S \to \mathbb{R} \) given by

\[
\hat{W}(s_1), \ldots, \hat{W}(s_n) \sim \mathcal{N}(0, \Sigma) \quad \text{and} \quad \mathbb{E}\{\hat{W}(s_1)\hat{W}(t_1) \cdots \hat{W}(s_n)\hat{W}(t_n)\} = \Sigma
\]

for any \( n \in \mathbb{N} \), \( a_1, \ldots, a_n \in \mathbb{R} \), and, \( s_1, \ldots, s_n \in S \), \( \sum_{ij} a_ia_jk(s_i, s_j) \geq 0 \).

We would also show that for any (continuous) non-negative definite function \( k \) on \( S \times S \), there exists a Gaussian \( \hat{W} \in C(S) \) with covariance function \( k \).

**Remark 1.** When \( \mathcal{F} = L_2(S, \nu) \), the evaluation functions \( \delta_s \) are no longer continuous, hence for a Gaussian \( \hat{W} \in \mathcal{F} \), it’s no longer meaningful to talk about normality of \( \hat{W}(s_1), \ldots, \hat{W}(s_n) \). However, for an orthonormal basis \( \{\phi_j(s) : j = 1, 2, \ldots \} \) of \( \mathcal{F} \), e.g., the Legendre polynomials if \( S = [0, 1] \) and \( \nu = \text{Lebesgue measure} \), the Fourier coefficients \( \hat{W}_j = \int \phi_j(s)W(s)\nu(ds) : j = 1, 2, \ldots \) \( \in \ell_2 \) satisfy the Gaussianity condition for a random element in \( \ell_2 \) (the space of square summable sequences). On \( \ell_2 \), coordinate projection is continuous, and hence, for any \( \{j_1, \ldots, j_n\} \subset \mathbb{N} \), \( X = (\hat{W}_{j_1}, \ldots, \hat{W}_{j_n})^T \) has a multivariate normal distribution. Note that \( W = \sum_j \hat{W}_j \phi_j \), and, if the latter series converges pointwise, then it’s meaningful to define \( W(s) = \sum_j \hat{W}_j \phi_j(s) \).
2.2 Gaussianity and Hilbert space

With every centered Gaussian element $W$ in a Banach space $(\mathcal{F}, \| \cdot \|)$, one can associate a Hilbert space¹ $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ where $\mathcal{H}$ is a continuously embedded² subset of $\mathcal{F}$. To see this, define the map $g: \mathcal{F}^* \to \mathcal{F}$ given by $L \mapsto gL := \mathbb{E}\{ (LW)^* \}$, with $(gL, gL') = \mathbb{E}\{ (LW)(L'W) \}$ and take $\mathcal{H}$ as the completion of $g(\mathcal{F}^*)$ under $\langle \cdot, \cdot \rangle$. It’s easy to see $\| f \| \leq \sigma \| f \|_{\mathcal{H}}$ where $\sigma^2 = \mathbb{E}\| W \|^2$, and hence, $\mathcal{H}$ is continuously embedded into $\mathcal{F}$. Because the Hilbert space norm is stronger than the Banach space norm, the completion of $g(\mathcal{F}^*)$ remains a subset of $\mathcal{F}$. When $W$ is non-degenerate, $\mathcal{H}$ is a dense subset of $\mathcal{F}$.

For this definition to be valid, one must be able to define $\mathbb{E}\{ (LW)^* \}$ and establish that if $h = gL_1 = gL'_1$ and $h' = gL_2 = gL'_2$, then $\mathbb{E}\{ (L_1W)(L_2W) \} = \mathbb{E}\{ (L'_1W)(L'_2W) \}$, i.e., the inner product is well defined. For $X$ a random element of a Banach space $(\mathcal{F}, \| \cdot \|)$, $\mathbb{E}(X)$ is defined as the unique element $f \in \mathcal{F}$ such that $Lf = \mathbb{E}(LX)$ for every $L \in \mathcal{F}^*$. Such a unique element exists as long as $\mathbb{E}\| X \| < \infty$. For our case, $X = (LW)W$ for some $L \in \mathcal{F}^*$, and hence $\| X \| = |LW|\| W \| \leq \| L \| \| W \|^2$. So we need $\mathbb{E}\| W \|^2 < \infty$. This is true for any centered Gaussian element $W$ in a separable Banach space, but we omit a proof³. Also notice that if $gL = gL'$, then with $\Delta = L - L' \in \mathcal{F}^*$, $0 = \Delta(g\Delta) = \mathbb{E}(\Delta W)^2$, hence $\Delta W = 0$ with probability one. Therefore, if $gL_1 = gL'_1$ and $gL_2 = gL'_2$,

$$\mathbb{E}\{ (L_1W)(L_2W) \} - \mathbb{E}\{ (L'_1W)(L'_2W) \} = \mathbb{E}\{ (\Delta_1W)(L_2W) \} + \mathbb{E}\{ (L'_1W)(\Delta_2W) \} = 0$$

where $\Delta_1 = L_1 - L'_1$ and $\Delta_2 = L_2 - L'_2$.

For any $h = gL \in \mathcal{H}$, define the scalar random variable $U(h) = LW$. This is well defined because, as shown above, if $h = gL = gL'$ then $(L - L')W$ is zero with probability one. Therefore we can associate with $W$ the stochastic process $U = (U(h)) : h \in \mathcal{H})$. For any finite set $\{ h_1, \ldots, h_n \} \subset \mathcal{H}$, the random vector $(U(h_1), \ldots, U(h_n))$ has an $n$-dimensional normal distribution with mean zero and variance-covariance matrix determined by: $\mathbb{E}\{ U(h_i)U(h_j) \} = \langle h_i, h_j \rangle$.

If $(\mathcal{F}, \| \cdot \|)$ is separable, it can be shown that $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is also separable⁴. Such an $\mathcal{H}$ admits an at most countable orthonormal basis $\{ \hat{h}_n : n \geq 1 \} \subset \mathcal{H}$. The following result is fundamental (see van der Vaart and van Zanten, 2008, Theorem 4.3)

**Theorem 1.** The random variables $U(\hat{h}_1), U(\hat{h}_2), \ldots$ are IID standard normal variables and

$$W = \sum_n U(\hat{h}_n)\hat{h}_n$$

where the series converges in the Banach space norm $\| \cdot \|$, almost surely.

---

¹Known as the Cameron-Martin space of $W$ in $\mathcal{F}$.

²That is $\mathcal{H} \subset \mathcal{F}$ and for some constant $C$, $\| h \| \leq C\| h \|_{\mathcal{H}}$ for every $h \in \mathcal{H}$.

³Fernique’s Theorem says that for any centered Gaussian $W$ in a separable Banach space $(\mathcal{F}, \| \cdot \|)$, $\mathbb{E}\exp\{ \| W \|^2/(18R^2) \} \leq \exp\{ c^2/(18R^2) \}$ where $R := \inf\{ r : \mathbb{P}(\| W \| \leq r) \geq 0.9 \}$; see Stroock (2010) Theorem 8.2.1.

⁴This is not a trivial result and generally is not guaranteed for any Hilbert space that is continuously embedded into a separable Banach space. But it holds for the Cameron-Martin space $\mathcal{H}$ we have here.
Two important consequences of this result are: (1) any Gaussian random element \( W \) in a separable Banach space \( (\mathcal{F}, \|\cdot\|) \) is completely determined by the Hilbert space \( (\mathcal{H}, \langle \cdot, \cdot \rangle) \) associated with it, and, (2) \( W \) can be expanded in a series \( W = \sum_n Z_n \hat{h}_n \) for any orthonormal basis of \( (\mathcal{H}, \langle \cdot, \cdot \rangle) \), where the series converges in the norm of the Banach space. In particular, the support of \( W \) (the smallest closed subset of \( \mathcal{F} \) that receives probability 1 under the law of \( W \)) is precisely the closure of \( \mathcal{H} \) in \( \mathcal{F} \) in the Banach space norm. Interestingly, \( \mathcal{H} \) consists of precisely the elements \( \sum_n a_n \hat{h}_n \) for which \( \sum_n a_n^2 < \infty \). But since \( U(\hat{h}_n) \) are independent standard normals, \( \sum_n U(\hat{h}_n)^2 = \infty \) with probability 1. And hence \( \Pr(W \in \mathcal{H}) = 0 \) even though \( \Pr(W \in \mathcal{H}) = 1 \).

One final remark is that even though the Cameron-Martin space \( (\mathcal{H}, \langle \cdot, \cdot \rangle) \) is unique for a Banach space valued centered, non-degenerate Gaussian element \( W \), the embedding Banach space need not be unique. That is one can find another Banach space \( \mathcal{F}' \) which in turn is a continuously embedded dense subset of \( \mathcal{F} \) which in turn is a continuously embedded dense subset of \( \mathcal{F} \). For example when a centered, non-degenerate Gaussian element of \( (C(S), \|\cdot\|_\infty) \) may have a Cameron-Martin space that could be embedded into a Hölder space \( (C^{\beta}(S), \|\cdot\|_\beta) \) for some \( \beta > 0 \).

### 2.3 Reproducing Kernel Hilbert Space

A Hilbert space \( (\mathcal{H}, \langle \cdot, \cdot \rangle) \) with \( \mathcal{H} \subset \mathbb{R}^S \) is said to be a reproducing kernel Hilbert space (RKHS) if \( \delta_s \) is continuous for every \( s \in S \) with respect to the Hilbert space norm \( \|h\|_\mathcal{H} = \langle h, h \rangle^{1/2} \). Note that the definition does not say anything about a ‘reproducing kernel’ – that’s coming later. Also, continuity is demanded with respect to the Hilbert space norm, not the original Banach space norm.

For the Hilbert space \( (\mathcal{H}, \langle \cdot, \cdot \rangle) \) associated with a Gaussian \( W \in C(S) \), for any \( h = gL, h' = gL' \in \mathcal{H} \) and any \( s \in S \),

\[
|\delta_s h - \delta_s h'| = |\mathbb{E}\{(LW)(W(s))\} - \mathbb{E}\{(L'W)(W(s))\}|
\leq [\mathbb{E}\{(W(s))^2\}]^{1/2}[\mathbb{E}\{(L - L')(W)\}]^{1/2} \quad [\text{by Cauchy-Schwartz}]
= [\mathbb{E}\{(W(s))^2\}]^{1/2}\|h - h'\|_\mathcal{H}.
\]

Therefore \( (\mathcal{H}, \langle \cdot, \cdot \rangle) \) is an RKHS.

Now we define a reproducing kernel. A function \( k : S \times S \to \mathbb{R} \) is called a reproducing kernel of \( (\mathcal{H}, \langle \cdot, \cdot \rangle) \) if

1. \( k(s, \cdot) \in \mathcal{H} \) for every \( s \in S \), where \( k(s, \cdot) : t \mapsto k(s, t) \), and,\footnote{For a multi-index \( n = (n_1, \ldots, n_d) \in \{0, 1, 2, \ldots\}^d \) let \( D^n \) denote the differential operator \( \partial^{n_1}/\partial x_1^{n_1} \cdots \partial x_d^{n_d} \), where \( |n| = n_1 + \cdots + n_d \). For any positive real \( r \) let \( \lfloor r \rfloor \) denote the largest integer strictly smaller than \( r \). For any \( \beta > 0 \), the Hölder space \( C^{\beta}(S) \) of order \( \beta \) is defined as the space of all functions \( f \) with continuous derivatives \( D^n f \) for all orders \( n \) with \( |n| \leq \lfloor \beta \rfloor \) and for any \( n \) with \( |n| = \lfloor \beta \rfloor \), \( \sup_{x \neq y \in S} |D^n f(x) - D^n f(y)|/\|x - y\|^{\beta - |\beta|} < \infty \). This space becomes a Banach space with norm \( \|f\|_\beta = \max_{\|n\| \leq \lfloor \beta \rfloor} \|D^n f\|_\infty + \max_{|n| = \lfloor \beta \rfloor} \sup_{x \neq y \in S} |D^n f(x) - D^n f(y)|/\|x - y\|^{\beta - |\beta|} \). This space is a continuously embedded dense subspace of \( C(S) \) equipped with \( \|\cdot\|_\infty \).}
2. \( \langle f, k(s, \cdot) \rangle = f(s) \) for every \( s \in S, f \in \mathcal{H} \).

If \( k \) is the covariance function of a centered Gaussian \( W \in C(S) \), then \( k \) satisfies both conditions with respect to the associated Hilbert space \((\mathcal{H}, \langle \cdot, \cdot \rangle)\):

1. \( k(s, \cdot) = \mathbb{E}\{W(s)W\} = \mathbb{E}\{\delta_s(W)\} = g(\delta_s) \in \mathcal{H} \), and,

2. for any \( f = gL \in \mathcal{H} \),
   \[
   \langle gL, k(s, \cdot) \rangle = \mathbb{E}\{(LW)W(s)\} = \delta_s(\mathbb{E}\{(LW)\}) = f(s).
   \]

The reason RKHS are called RKHS is the following result: \( (\mathcal{H}, \langle \cdot, \cdot \rangle) \) is an RKHS if and only if it has a reproducing kernel, and, if a reproducing kernel exists then it is unique. Clearly for every centered Gaussian element \( W \in C(S) \), the associated Hilbert space is an RKHS with reproducing kernel given by the covariance function \( k(s, t) = \mathbb{E}\{W(s)W(t)\} \), which we know is a non-negative definite function on \( S \times S \). Conversely, for every positive definite function \( k : S \times S \to \mathbb{R} \), there exists a Hilbert space \((\mathcal{H}, \langle \cdot, \cdot \rangle)\) with reproducing kernel \( k \). This RKHS can be constructed by taking the completion of

\[
\mathcal{H}_0 = \{ h = \sum_{i=1}^{n} a_i k(s_i, \cdot) : n \in \mathbb{N}, a_1, \ldots, a_n \in \mathbb{R}, s_1, \ldots, s_n \in S \}
\]

under the inner product

\[
\langle \sum_{i=1}^{n} a_i k(s_i, \cdot), \sum_{j=1}^{m} b_j k(t_j, \cdot) \rangle := \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j k(s_i, t_j).
\]

When \( k \) is continuous, \( (\mathcal{H}, \langle \cdot, \cdot \rangle) \) can be shown to be separable. Let \( \{ \hat{h}_n : n \geq 1 \} \) be any orthonormal basis. Then,

\[
k(s, t) = \sum_{n=1}^{\infty} \langle k(s, \cdot), \hat{h}_n \rangle \hat{h}_n(t) = \sum_{n=1}^{\infty} \hat{h}_n(s) \hat{h}_n(t),
\]

where the last equality follows from the reproducing property of \( k \). Therefore the random element \( W = \sum_n Z_n \hat{h}_n \) where \( Z_n \overset{\text{iid}}{\sim} N(0, 1) \), is a centered Gaussian element of \( C(S) \) with covariance function \( k \).

### 3 Examples of covariance functions a.k.a kernels

The word kernel in ‘reproducing kernel’ has a significance. A function \( k : S \times S \to \mathbb{R} \) is called a kernel if there exists a Hilbert space \((\mathcal{H}, \langle \cdot, \cdot \rangle)\) and a map \( \phi : S \to \mathcal{H} \) such that \( k(s, t) = \langle \phi(s), \phi(t) \rangle \). A reproducing kernel is a kernel since \( k(s, t) = \langle k(s, \cdot), k(t, \cdot) \rangle \). A kernel is clearly non-negative definite. Conversely, Any non-negative definite function is a kernel. Therefore, by our preceding discussion, for every kernel there exists a Gaussian process whose covariance function equals the kernel.
Non-negative constants. For $S \subset \mathbb{R}^d$, $k(s, t) \equiv c$ for some $c \geq 0$ is a (trivial) kernel since its the covariance function of the constant stochastic process $W(s) \equiv Z$ where $Z \sim N(0, c)$.

Linear kernel. For $S \subset \mathbb{R}^d$, $k(s, t) = s^T t$ is clearly a kernel by definition.

Product of kernels. If $S = S_1 \times S_2$ and $k_1 : S_1 \times S_1 \rightarrow \mathbb{R}$, $k_2 : S_2 \times S_2 \rightarrow \mathbb{R}$ are kernels then $k = k_1 \otimes k_2 : S \times S \rightarrow \mathbb{R}$ given by $k((s_1, s_2), (t_2, t_2)) = k_1(s_1, t_1)k_2(s_2, t_2)$ is a kernel. To see this notice that there are Gaussian process $(W_1(s_1) : s_1 \in S_1)$ and $(W_2(s_2) : s_2 \in S_2)$ with covariance functions $k_1$ and $k_2$ respectively. We can take $W_1$ and $W_2$ to be independent. Define the (non-Gaussian) stochastic process $W(s_1, s_2) = W_1(s_1)W_2(s_2)$ which is easily seen to have covariance function $k$, hence $k$ must be positive definite and hence it must be a kernel.

A special case is when $S_1 = S_2 = S$, then $k(s, t) = k_1(s, t)k_2(s, t)$ is a kernel on $S \times S$.

Linear combination of kernels with non-negative coefficients. If $k_1, k_2$ are two kernels on $S \times S$ then so is $k = ak_1 + bk_2$ for any non-negative reals $a$ and $b$. This is trivial to see since $k$ is the covariance function of $W(s) = \sqrt{a}W_1(s) + \sqrt{b}W_2(s)$ where $W_1, W_2$ are independent centered Gaussian processes with covariance functions $k_1, k_2$ respectively.

Polynomial kernels. By combining the above two results, for any kernel $k : S \times S \rightarrow \mathbb{R}$, and a polynomial $p(x) = a_m x^m + \cdots + a_1 x + a_0$ on $(0, \infty)$ with non-negative coefficients, the function $k(s, t) = p(k(s, t))$ defines a kernel. By extension, any function $f : (0, \infty) \rightarrow \mathbb{R}$ that admits a power series expansion on $(0, \sup_x k(s, t))$ gives a new kernel $\tilde{k}(s, t) := f(k(s, t))$. An example of such a function is $f(x) = e^x$, and, hence, $k(s, t) = e^{s^T t}$ is a kernel on $S \times S \rightarrow \mathbb{R}$ for any $S \subset \mathbb{R}^d$.

Square exponential kernel. $k(s, t) = e^{-(s-t)^2/2}$ is a kernel on $\mathbb{R} \times \mathbb{R}$ since one can write: $k(s, t) = e^{-s^2/2}e^{-t^2/2}e^{st}$ which is the covariance function of $W(s) = e^{-s^2/2}\tilde{W}(s)$ where $\tilde{W}(s)$ is a Gaussian process with covariance kernel $e^{st}$. By the product rule, $k(s, t) = \exp\{-\|s-t\|^2/2\}$ is a kernel on $\mathbb{R}^d \times \mathbb{R}^d$.

Scaling. If $k(s, t)$ is a kernel on $S \times S$ for $S \subset \mathbb{R}^d$ then for any non-zero $d \times q$ matrix $A$ and any $S' \subset \mathbb{R}^d$ with $AS' \subset S$, $k_A(s', t') = k(As', At')$ is a kernel on $S' \times S'$. By applying to the square exponential kernel above, $k(s, t) = \exp\{-a^2\|s-t\|^2\}$ is a kernel for any $a > 0$. Also, for any $d \times d$ matrix $A$, $k(s, t) = \exp\{-\|A(s-t)\|^2\}$ is a kernel. By choosing, $A = \text{diag}(a_1, \ldots, a_d)$, $a_j \geq 0$, one gets the anisotropic square-exponential kernel $k(s, t) = \exp\{-\sum_{j=1}^d a_j^2(s_j - t_j)^2\}$ on $\mathbb{R}^d \times \mathbb{R}^d$. 

Polynomial splines. For any $m \in \mathbb{N}$ one can define $k : [0, 1] \times [0, 1] \to \mathbb{R}$ as:

$$k(s, t) = \sum_{i=1}^{m} \frac{s^{i-1} t^{i-1}}{(i-1)!^2} + \int_{0}^{1} \frac{(s-u)^{m-1} (t-u)^{m-1}}{((m-1)!)^2} du$$

To see that this is a kernel, consider the Hilbert space $H = \mathbb{R}^m \times L^2[0, 1]$ equipped with the natural inner product $\langle (x, f), (y, g) \rangle = x^T y + \int_{0}^{1} f(u) g(u) du$. Clearly $k(s, t) = \langle \phi(s), \phi(t) \rangle$ where $\phi : [0, 1] \to \mathbb{R}^m \times L^2[0, 1]$ is given by

$$\phi(t) = \left( 1, t, \frac{t^2}{2}, \ldots, \frac{t^{m-1}}{(m-1)!}, u \mapsto \frac{(t-u)^{m-1}}{(m-1)!} \right),$$

where the last element is a function. The RKHS associated with this kernel is the space of all functions on $[0, 1]$ with $(m-1)$ absolutely continuous derivatives and an $m$-th derivative with a finite $L_2$ norm.

4 Sample path continuity and differentiability

Given a kernel $k : S \times S \to \mathbb{R}$ we know there exists a centered Gaussian process $W = (W(s) : s \in S)$ with covariance $k(s, t)$. What can we say about $W$ being an element of $C(S)$ with probability one, i.e, the sample paths of the process being continuous almost surely? Could we go further and say that $W$ has sample paths that are almost surely differentiable up to certain order? Can we say that $W$ is an element of $C^\beta(S)$ for some $\beta > 0$?

4.1 Continuity

A fairly long series of work by Dudley, Belyaev, Slepian, Ferniqué, Sudakov, Talagrand and others lead to a very sharp result about sample path continuity of a centered Gaussian process $W$ with covariance $k : S \times S \to \mathbb{R}$ defined on a compact set $S$. Define the so called canonical metric $\rho(s, t)$ on $S$ associated with $W$ as:

$$\rho(s, t) = \sqrt{\mathbb{E} \{ W(s) - W(t) \}^2} = \sqrt{k(s, s) + k(t, t) - 2k(s, t)}$$

which satisfies non-negativity, symmetry, and triangle inequality conditions of a metric, and, also satisfies identifiability if $W$ is non-degenerate. For any $\epsilon > 0$, let $N(\epsilon, S, \rho)$ denote the minimum number of $\rho$-balls with radius $\epsilon$ needed to cover $S$. This number is finite since $S$ is assumed compact (it remains compact with respect to $\rho$). It is called the $\epsilon$-covering number of $S$ under $\rho$, and log $N(\epsilon, S, \rho)$ is called the $\epsilon$-entropy of $S$ under $\rho$.

**Theorem 2.** If $\int_{0}^{\infty} \sqrt{\log N(\epsilon, S, \rho)} d\epsilon < \infty$ then $W$ is a random element of $C(S)$. 
A more precise statement would be to say that a version of $W$ exists with almost surely continuous sample paths\textsuperscript{6}. Since $S$ is compact, log $N(\epsilon, S, \rho) = 0$ for all $\epsilon > \text{diam}(S) := \sup_{s, t \in S} \rho(s, t)$. Therefore the issue of finiteness is at the lower end of the integral. The entropy needs to grow slowly as $\epsilon$ shrinks to zero.

**Example (Lipschitz continuous $k$).** Suppose that $k : S \times S \to \mathbb{R}$ is Lipschitz continuous on a compact $S \subset \mathbb{R}^d$, i.e.,

$$M := \sup_{(s, t) \neq (s', t') \in S \times S} \frac{|k(s, t) - k(s', t')|}{\|s - t\|} < \infty.$$  

Then $\rho^2(s, t) < 2M\|s - t\|$ for all $s, t \in S$. Therefore the $\epsilon$-entropy of $S$ in $\rho$ is smaller than or equal to the $\epsilon^2/(2M)$-entropy of $S$ in the Euclidean metric. But the latter entropy grows at the order of $2d \log(1/\epsilon)$ as $\epsilon \downarrow 0$, and hence the entropy condition of Theorem 2 holds. Therefore, Lipschitz continuity of $k$ implies sample path continuity of $W$. This covers a broad range of covariance kernels, including the Brownian motion covariance function $k(s, t) = \min(s, t)$ and the squares exponential kernel $k(s, t) = \exp\{-(\|s - t\|)^2\}$.

### 4.2 Hölder continuity

Although Lipschitz continuity of $k$ implies sample path continuity of $W$, it does not guarantee that the sample paths are Lipschitz continuous. In fact, the Brownian motion is known to have sample paths that are not Lipschitz with probability one. Under stronger condition one can establish stronger continuity properties. Recall that a function $f : S \to \mathbb{R}$ is called Hölder continuous\textsuperscript{7} with exponent $\alpha \in (0, 1]$ if $\sup_{s \neq t \in S} |f(s) - f(t)|/\|s - t\|^\alpha < \infty$. The famous Kolmogorov Continuity Theorem says that if for some $p, b, c > 0$,

$$\mathbb{E}\{|W(s) - W(t)|^p\} \leq c\|s - t\|^{1+b}, \text{ for all } s, t \in S,$$

then $W$ has sample paths that are Hölder continuous of exponent $\alpha$ for every $\alpha \in [0, b/p)$. It can be shown that sample paths of a Brownian motion are Hölder continuous of exponent $\alpha$ for all $\alpha < 1/2$.

### 4.3 Differentiability

Since differentiation is a linear operation, the very definition of Gaussianity says that the derivative of a Gaussian process is itself a Gaussian process. For this to be meaningful, however, one must ensure that the sample paths of the original process are

\textsuperscript{6}For a fairly accessible proof see Theorem 1 of \url{https://galton.uchicago.edu/~lalley/Courses/386/GaussianProcesses.pdf}.

\textsuperscript{7}this matches our definition of the Hölder space $C^\alpha(S)$.  

\hspace{10cm} 8
differentiable with probability one. As one would expect, path differentiability can be established from differentiability properties of the covariance kernel \( k \). For notational ease, let us begin with the one dimensional case.

**Theorem 3.** Let \( W \) is a centered, Gaussian process on a compact \( S \subset \mathbb{R}^1 \) with covariance kernel \( k : S \times S \rightarrow \mathbb{R} \). Suppose \( k(s,t) \) is Lipschitz continuous and twice continuously differentiable. If the mixed second partial derivative function

\[
r(s,t) := \frac{\partial^2}{\partial s \partial t} k(s,t), \; s,t \in S,
\]

is Lipschitz on \( S \times S \), then \( W \) is differentiable with probability one, and, the derivative process \( \dot{W}(s) \) is a Gaussian process with covariance function \( r(s,t) \).

A proof can be given as follows. Let \( T = \{(s,h) : s \in S, s + h \in S \} \). Define a new centered, Gaussian process \( Y(s,h) \) on \( D = T \setminus (S \times \{0\}) \) as

\[
Y(s,h) = \frac{W(s+h) - W(s)}{h}, \; s \in S, s + h \in S, h \neq 0,
\]

which, by the assumption on \( k \), can be extended to the closure of \( D \), which is \( T \). Equate \( \dot{W}(s) = Y(s,0) \).

In the case of \( S \subset \mathbb{R}^d \) with \( d \) possibly larger than 1, the derivative process \( \nabla W(s) = (\frac{\partial}{\partial s_1} W(s), \ldots, \frac{\partial}{\partial s_d} W(s)) \), if it exists, is \( d \)-dimensional. Theorem 3 then needs to be modified with

\[
r(s,t) = \nabla^2 k(s,t) = \left( \frac{\partial^2}{\partial s_i \partial t_j} k(s,t) \right)_{i,j=1}^d
\]

being the (matrix-valued) covariance function\(^8\) of \( \nabla W \).

## 5 Gaussian process regression

In many situations we require borrowing information across observation units based on covariate information. This is manifest in the standard regression / supervised learning setting where \( n \) observations \((Y_i, x_i), \; i = 1, \ldots, n\), on an outcome of interest \( Y \) and covariate \( X \) are to be used to predict \( Y \) at a new \( X = x^* \), with the understanding that data units with \( x_i \) more similar to \( x^* \) should have a bigger influence on prediction at \( x^* \). Borrowing may also be required for latent variables/parameters within a larger hierarchical model. Going back to our mixed effects model: \( Y_{ij} = \beta^T x_{ij} + b_i^T z_{ij} + \epsilon_{ij} \), we may require \( b_i \)'s to be more similar for group indexes \( i \) with similar covariate information \( s_i \) which may comprise of spatial location, group level socio-economic characteristics,

---

\(^8\)Any function \( f : S \rightarrow \mathbb{R}^d \) can also be viewed as a function \( \tilde{f} : \{1, \ldots, d\} \times S \rightarrow \mathbb{R} \), with \( \tilde{f}(i,s) = f_i(s) \). In my opinion, vector valued stochastic processes are best viewed through such extension mappings.
etc. For simplicity of exposition, we present the following models in the regression setting, and later extend them to more general settings as needed.

A flexible model that allows borrowing of information across units in the regression setting is:

$$Y_i = f(x_i) + \epsilon_i$$

where the unknown regression function $f$ is restricted to belong to some space of continuous or even differentiable functions. For simplicity, we work with continuous, unbounded outcome $Y$ and take $\epsilon_i \sim N(0, \sigma^2)$. Other response models could be constructed by the use of suitable link functions, analogous to GLM formulations. The conditional distribution of $Y$ given $X = x^*$ and $f$, is then $N(f(x^*), \sigma^2)$. When data $(Y_i, x_i), i = 1, \ldots, n$ are observed, they carry information about $f$ at $x \in \{x_1, \ldots, x_n\}$. This information percolates to $f(x^*)$ through continuity and smoothness assumptions on $f$, with information from nearby $f(x_i)$ exerting a bigger influence.

For Bayesian inference and prediction, we need a prior distribution on $f$. From what we have discussed earlier, a practical choice is a Gaussian process distribution $GP(m, C)$, with $C$ chosen so that the sample paths of $GP(m, C)$ have desirable continuity and smoothness properties. We will see later that one rarely gets all the necessary flexibility by using a single covariance function $C$. Instead, more efficient learning from data usually takes place when a family of covariance function $C(\mid \theta)$ is considered, indexed by a low dimensional parameter $\theta \in \Theta$. A widely used example is the squared exponential covariance

$$C(s, t \mid \tau, \psi) = \tau^2 \exp\{-\psi^2 \|s - t\|^2\},$$

(indexed by two positive scalars $\tau$ and $\psi$. In the square exponential case, the parameter $\psi$, often called the correlation-range parameter, controls the relative modulus of continuity of the sample paths. It is also possible to allow different moduli of continuity along different elements of the covariate space:

$$C(s, t \mid \tau, \psi) = \tau^2 \exp\{-\sum_{j=1}^{p} \psi_j^2 (s_j - t_j)^2\},$$

with $\psi = (\psi_1, \ldots, \psi_p) \in (0, \infty)^p$. This is often referred to as the anisotropic square exponential covariance. From now on, fix $m \equiv 0$.

5.1 Conjugacy wrt Gaussian likelihood

For the Gaussian error regression model, the (conditional) posterior distribution of $f$ given $D_n = \{(Y_i = y_i, x_i) : 1 \leq i \leq n\}$ (and $\theta, \sigma^2$) is again a Gaussian process [HW 2]
with\(^9\)

\[
\mathbb{E}[f(x)|D_n] = c_n^T(x)(C_{nn} + \sigma^2I_n)^{-1}y
\]

\[
\text{Cov}[f(s), f(t)|D_n] = C(s, t) - c_n^T(s)(C_{nn} + \sigma^2I_n)^{-1}c_n(t)
\]

where \(y = (y_1, \ldots, y_T)^T\), \(c_n(x) = (C(x, x_1), \ldots, C(x, x_n))^T\) and \(C_{nn}\) is the \(n \times n\) covariance matrix with \((i, j)\)-th entry \(C(x_i, x_j)\). Also, it is straightforward to see that we can write down the marginal pdf of \(Y = (Y_1, \ldots, Y_n)\) by integrating out \(f\) as

\[
p(y|\theta, \sigma^2) = N_n(0, C_{nn} + \sigma^2I_n)
\]

and hence the marginal log-likelihood of \((\sigma^2, \theta)\) may be calculated as:

\[
\log p(y|\sigma^2, \theta) = -\frac{1}{2} \log \det(C_{nn} + \sigma^2I_n) - \frac{1}{2} y^T(C_{nn} + \sigma^2I_n)^{-1}y.
\]

The plots in Figure 1 show the pointwise posterior means and credible bands (mean \(\mp\) 2 standard deviation) of \(f\) based on \(n = 10\) observations that were generated from the GP model as above with \(\sigma = 0.1\), \(\tau = 1\) and \(\psi = 10\). In the plots, the posterior mean and variance of \(f\) were calculated using the known values of \(\sigma\) and \(\tau\), but 4 different values of \(\psi \in \{1, 5, 10, 20\}\) were tried. Also reported are the resulting log marginal likelihood.

### 5.2 A full model specification

Toward a full specification of the GP regression prior, we use the anisotropic covariance kernel (2), but reparametrize \(\tau = \sigma \rho\). The parameter \(\rho\) can be interpreted as signal-to-noise-ratio and the quantity \(R^2 = \rho^2/(1 + \rho^2)\) is akin to R-square model-fit statistic in ordinary regression. We specify a \(Be(h_1, h_2)\) prior on \(R^2\) and work with the induced prior on \(\rho\). For a “default specification” I will work with \(h_1 = 1.36, h_2 = 1\) which ensure \(R^2\) has a prior median of 60\% and is between 6\% and 98\% with prior probability 0.95.

The reparametrization also helps to specify a conjugate inverse-gamma prior on \(\sigma^2\), which can then be analytically integrated out of the model during model fitting. I will work with \(\sigma^{-2} \sim Ga(r/2, r\sigma_0^2/2)\) with \(\sigma_0^2\) chosen (depending on \(r\)) so that \(\Pr(\sigma^2 \leq 1) = 0.95\). This is done assuming the response data is standardized to have mean zero and variance 1. So our prior on \(\sigma^2\) encodes 95\% confidence that we can do better than predicting \(Y\) to be zero at every predictor value. As a default choice, I will work with \(r = 2.5\) which gives \(\Pr(\sigma \leq 1/3) = 1/2\).

Finally, for a prior specification on \(\psi = (\psi_1, \ldots, \psi_p)\), we take \(\psi_j\)'s to be independently distributed such that \(\Pr(\psi_j^2 \leq s) = \Pr(T \leq s)^{1/p}\) where \(T \sim Ga(a_1, a_2)\). With

\(^9\)An interesting special case happens in the “noise-free” case when \(\sigma = 0\): the mean function passes through the points \((x_i, y_i)\) and the variance function at each observed \(x_i\) is exactly zero. Therefore, in the noise-free case, the sample paths from the posterior GP pass through \((x_i, y_i)\) with probability 1, i.e., they interpolate the observed data.
Figure 1: Pointwise posterior mean and credible band (mean ± 2×SD) of $f$ for toy data simulated from the mean zero square exponential GP regression model (1) with $\psi = 10$, $\sigma = 0.1$ and $\tau = 1$. The dashed gray lines show the same under the prior. Posterior computation were done using the known values for $\sigma$ and $\tau$, but 4 different values of $\psi$ were tried. Plot titles report the corresponding log-marginal likelihood scores.
this choice, \( \max_j \psi_j^2 \sim Ga(a_1, a_2) \). Notice that the quantity \( c(\Delta) := \exp(-\Delta^2 \max_j \psi_j^2) \) precisely equals

\[
c(\Delta) = \inf_{u: \|u\|=1} \text{Cor}(f(x), f(x + \Delta u)).
\]

I will work with \( a_1 = 5, a_2 = 0.3 \), so that, \textit{a priori}, \( \Pr(0.71 \leq c(\Delta) \leq 0.95) = 0.95 \).

For likelihood evaluation, we can clearly integrate out \( \sigma^2 \) from the model and write:

\[
\log p(y|\theta) = -\frac{1}{2} \log \det (\rho^2 K_{nn} + I_n) - \frac{n+r}{2} \log \left( 1 + \frac{y^T(\rho^2 K_{nn} + I_n)^{-1}y}{r\sigma_0^2} \right)
\]

where, now, \( K_{nn} \) denotes the \( n \times n \) correlation matrix \( \rho^{-2}C_{nn} \).

### 6 Low rank approximation

A computational bottleneck in implementing GP regression is the \( O(n^3) \) computational complexity associated with inverting or factorizing the \( n \times n \) matrix \( C_{nn} + \sigma^2 I_n \). However, when the underlying covariance function \( C(s,t) \) is smooth, the covariance matrix \( C_{nn} \) can be expected to be dense and rank-deficient, i.e., its eigenvalues are likely to decay rather rapidly; see Figure 2. So, a reasonable rank-\( m \) approximation to \( C_{nn} \) maybe obtained by switching off its last \( n-m \) eigenvalues to zero. In fact, this zeroing of eigenvalues gives the optimal rank-\( m \) approximation to \( C_{nn} \) with respect to the \( \ell_2 \) operator norm. However, computing each eigenvalue requires \( O(n^3) \) arithmetic operations, and hence such a strategy fails to reduce any computational burden. An alternative is to use an incomplete, rank-\( m \) Cholesky factorization of \( C_{nn} \), which could be computed in \( O(nm^2) \) time.

It turns out that a rank-\( m \) incomplete Cholesky factorization can be interpreted as finite rank stochastic process approximation to the original infinite rank GP. Below I review the probability theory that makes this connection clear and precise. Intuitively, for a GP \( f \) over \( \mathcal{X} \) whose sample paths are smooth (at least continuous), one could track the process only at a set of knots \( t^*_1, \ldots, t^*_m \in \mathcal{X} \) and reconstruct approximate versions of the sample paths by “connecting the dots”. If the knots are space filling, then the reconstructed sample paths should be close to the original sample paths. The incomplete Cholesky factorization corresponds to a special way of connecting the dots, often referred to as the \textit{predictive process approximation} or the \textit{projected process approximation}.

#### 6.1 Finite rank predictive process approximation to a GP

Let \( f \sim \text{GP}(0, C) \) over \( \mathcal{X} \). Fix a set of points, referred to as \textit{knots} hereafter, \( \{t^*_1, \ldots, t^*_m\} \subset \mathcal{X} \) and write \( f = \tilde{f} + r \), where the \textit{predictive process} \( \tilde{f} \) is defined as

\[
\tilde{f}(x) = \mathbb{E}\{f(x) \mid f(t^*_1), \ldots, f(t^*_m)\}, \quad x \in \mathcal{X}.
\]
Then \( \tilde{f} \) defines a stochastic process on \( \mathcal{X} \). It is finite rank, because one can write

\[
\tilde{f}(x) = \sum_{i=1}^{m} \alpha_i C(t_i^*, x), \quad x \in \mathcal{X},
\]

with the coefficient vector \( \alpha = (\alpha_1, \ldots, \alpha_m)^T \sim N(0, C_{ss}^{-1}) \), where \( C_{ss} = ((C(t_i^*, t_j^*)) \). Clearly \( \tilde{f}(t_i^*) = f(t_i^*) \) at every \( i = 1, \ldots, m \), and hence \( \tilde{f} \) does connect the dots between \( f \) evaluated at the knots. However, the connecting has been done in a special way, using distributional properties of \( f \) itself. By replacing \( f \) with \( \tilde{f} \) in the statistical model, one now deals with the covariance matrix \( \tilde{C}_{nn} \) of \( \tilde{f}_n = (\tilde{f}(x_1), \ldots, \tilde{f}(x_n))^T \), which, due to the rank-\( m \) property of \( \tilde{f} \), can be factorized in \( O(nm^2) \) time.

Let \( \delta = \sup_{t \in \mathcal{X}} \min_{1 \leq i \leq m} \rho(t, t_i^*) \) denote the mesh size of the knots in the canonical metric \( \rho(t, s) = \left[ E\{f(t) - f(s)\}^2 \right]^{1/2} \) of the GP \( f \sim \text{GP}(0, C) \). When \( C(t, s) \) is continuous, \( \delta \) can be made arbitrarily small by packing \( \mathcal{X} \) with sufficiently many, well placed knots. But, as \( \delta \) tends to 0, so does \( \kappa^2 = \sup_{t \in \mathcal{X}} \text{Var}\{r(t)\} \). This is because for any \( t \in \mathcal{X} \), and any \( i \in \{1, \ldots, m\} \), by the independence of \( \tilde{f} \) and \( r \),

\[
\text{Var}\{r(t)\} = \text{Var}\{f(t)\} - \text{Var}\{f(t_i^*)\}
= E[\text{Var}\{f(t)\} | f(t), \ldots, f(t_i^*)]\]
= E[\text{Var}\{f(t) - f(t_i^*)\} | f(t), \ldots, f(t_i^*)]\]
\leq \text{Var}\{f(t) - f(t_i^*)\} = \rho^2(t, t_i^*),
\]

and hence \( \kappa \leq \delta \). That \( \kappa \) can be made arbitrarily small is good news, because it plays a key role in providing probabilistic bounds on the residual process \( r \).

**Theorem 4.** Assume \( \mathcal{X} \subset \mathbb{R}^p \).

(i) If \( \mathcal{X} \subset [a, b]^p \) and there is a finite constant \( c > 0 \) such that \( \text{Var}\{f(s) - f(t)\} \leq \lambda^2 \| s - t \|^2 \), \( s, t \in \mathcal{X} \) then

\[
P \left( \sup_{x \in \mathcal{X}} |r(x)| > \epsilon \right) \leq 5 \exp \left( -\frac{\epsilon^2}{B^2 \kappa} \right), \quad \forall \epsilon > 0 \quad (5)
\]
with $B = 72 \sqrt{2p^{3/2} \chi(b-a)}$ and \( \kappa^2 = \sup_{x \in \mathcal{X}} \text{Var}\{r(x)\}$.

(ii) For any finite subset $S \subset \mathcal{X}$

$$P \left( \sup_{x \in S} |r(x)| > \epsilon \right) \leq \exp \left\{ -\frac{\epsilon^2}{4\kappa^2_S (2 + \log |S|)} \right\}, \forall \epsilon > 0$$

(6)

where $|S|$ denotes the size of $S$ and $\kappa^2_S = \sup_{x \in S} \text{Var}\{r(x)\}$.

Note that the constant $B$ does not depend on the number or locations of the knots, it only depends on the dimensionality and size of $\mathcal{X}$ as well as smoothness properties of the covariance function $f$. It is possible to replace $\kappa$ in (5) with $\kappa^{2(1-\eta)}$ for any arbitrary $\eta \in (0, 1)$, but with a different constant $B$. While (5) provides an accuracy bound over the entire domain $\mathcal{X}$, the bound in (6) over a finite subset maybe of more practical value.

For Gaussian process regression models with additive Gaussian noise, a common modification (Finley et al., 2009) of predictive process approximation is to replace $f$ with the process $\hat{f} = \tilde{f} + r^*$ where $r^*$ is a zero mean Gaussian process, independent of $\tilde{f}$ and $r$, satisfying,

$$\text{Cov}\{r^*(t), r^*(s)\} = \begin{cases} \text{Cov}\{r(t), r(s)\} = \text{Var}\{r(t)\} & \text{if } t = s \\ 0 & \text{if } t \neq s. \end{cases}$$

The addition of $r^*$ gives $\hat{f}$ the same pointwise mean and variance as those of $f$, without adding to the computational cost. The residual process is now given by $\hat{r} = f - \hat{f} = r - r^*$ whose variance equals $2\text{Var}\{r(t)\}$ because of independence between $r$ and $r^*$. Because $\hat{r}$ is almost surely discontinuous, the bound in (5) does not apply. But (6) continues to hold with $\kappa^2_S$ replaced by $\hat{\kappa}^2_S = 2\kappa^2_S$.

7 Adaptive finite rank approximation

7.1 Incomplete Cholesky factorization

Let $S = \{x_1, \ldots, x_n\}$. With a slight abuse of notation, let $C = ((c_{ij}))$ denote the $n \times n$ covariance matrix of the Gaussian vector $f_n = (f(x_1), \ldots, f(x_n))^T$. A Cholesky factor $L = ((\ell_{ij}))$ of $C$, with $L$ being a $n \times n$ lower triangular matrix with non-negative diagonal elements and satisfying $C = LL^T$ must also satisfy the system of equations:

$$\sum_{j=1}^{i} \ell_{ij} L_{sj} = C_{si}, \quad i = 1, \ldots, n,$$

where the notation $A_{sj}$ is used to denote the $j$-th column of a matrix $A$. Therefore $L$ can be constructed one column at a time, from left to right, by using the following
formulae to construct the $i$-th column given $L_{sj}$, $j = 1, \ldots, i - 1$:
\[
\ell_{ii} = \sqrt{c_{ii} - \sum_{j<i} \ell_{ij}^2}, \quad \ell_{ki} = \frac{c_{ki} - \sum_{j<i} \ell_{ij} \ell_{kj}}{\ell_{ii}}, \quad k = i + 1, \ldots, n,
\]
(7)

with $\ell_{ki}$, $k < i$ set to zero. This is repeated for $i = 1, \ldots, n$, requiring a total of $O(n^3)$ computation time.

For an $m \in \{1, \ldots, n\}$, an approximation $\tilde{L}$ to $L$ obtains in $O(nm^2)$ time by an incomplete application of the above recursion stopping exactly at column $m$, and the remaining entries set to zero. This $\tilde{L}$ is precisely the Cholesky factor of the rank-$m$ predictive process $\tilde{f} = \mathbb{E}\{f|f(s_1), \ldots, f(s_m)\}$. Moreover, the resulting residual process $r = f - \tilde{f}$ satisfies $\text{Var}\{r(s_i)\} = \tilde{d}_i$, $1 \leq i \leq n$ where,
\[
\tilde{d}_i = 0, i \leq m; \quad \tilde{d}_i = \sqrt{c_{ii} - \sum_{j\leq m} \ell_{ij}^2}, i = m + 1, \ldots, n,
\]
and hence $\kappa_S = \max_{x \in S} \text{Var}\{r(x)\}^{1/2} = \max_{i>m} \tilde{d}_i$. Notice that the vector $\tilde{d}$ is computed in an additional $O(nm)$ time.

Therefore, the above incomplete Cholesky factorization produces a Gaussian predictive process approximation, with readily available error bounds, provided we are happy to choose the knots from the set $S$. The restriction to $S$ appears reasonable for most applications with the additional burden on the modeler to identify $S$ carefully. For example, in a Gaussian process regression model with additive noise, it is sufficient to take $S$ to be the training set of covariate values, if only posterior predictive mean and variances are needed at test cases. But if posterior predictive covariance between two test cases, or a test and a training case is required, then $S$ should include these test cases as well.

### 7.2 Pivoting and dynamic stopping

Notice that $\kappa_S$ decreases monotonically as the recursion progresses and more columns are constructed. This can be seen directly by looking at the formulas above: each $\tilde{d}_i$ decreases as $m$ is incremented. It can also be inferred from the GP analogy that $\text{Var}\{f(x)|f(s_1), \ldots, f(s_m)\} \leq \text{Var}\{f(x)|f(x_1), \ldots, f(x_m), f(x_{m+1})\}$. So a stopping rule can be devised by pre-specifying a tolerance level $\kappa_{\text{tol}}$ and stopping at $m$ which brings $\kappa_S$ below this tolerance.

Notice that the selected knots are always the first $m$ elements of the sequence in which $\{x_1, \ldots, x_n\}$ were originally arranged. This might not be an optimal choice if it was desired to reach the tolerance with smallest possible rank $m$ [which should be the goal given the computing time increases quadratically in $m$]. A greedy algorithm that looks for an “optimal” reordering of the original sequence is as follows. If current $\kappa_S$ does not meet the tolerance, first swap $x_{m+1}$ with $x_{i^*}$ where $i^* = \arg\max_{i>m} \tilde{d}_i$ and then increment $m$ and construct the new column of $L$. 

16
The dynamic swapping is a common feature, known as pivoting, of all leading software packages for Cholesky factorization. If run until \( m = n \), pivoting produces a permutation \( \pi = (\pi_1, \ldots, \pi_n) \) of \((1, \ldots, n)\) and a lower triangular matrix \( L \) with non-negative diagonals such that \( P_{\pi}C_{\pi}^T = LL^T \) where \( P_{\pi} \) is the \( n \times n \) permutation matrix associated with \( \pi \). The proposal above simply adds to this pivoted Cholesky factorization a dynamic, tolerance based stopping. The resulting \( \tilde{L} \) gives the Cholesky factor of the covariance matrix of \( \tilde{f}_n = (\tilde{f}(x_1), \ldots, \tilde{f}(x_n)) \) where \( \tilde{f} \) is the Gaussian predictive process associated with the knots \( x_{\pi_1}, \ldots, x_{\pi_m} \). The process \( \tilde{f} \) comes with the error bound (6) with \( \kappa_{\text{tol}} \) replacing \( \kappa_S \). The additional computing time needed for pivoting is only \( O(nm) \), a small fraction of the computing time \( O(nm^2) \) needed to get the elements \( \tilde{L} \) if \( \pi \) was precomputed.

7.3 Low-rank model fitting

- Data: standardized response vector \( y \in \mathbb{R}^n \), standardized design matrix \( x \in \mathbb{R}^{n \times p} \).
- Model: \( y = f + \epsilon \) where \( \epsilon \sim N(0, \sigma^2 I_n) \).
- Prior: \((f, \sigma^2) \sim N(0, \rho^2\sigma^2 K) \times IG(r/2, r\sigma^2_0/2)\) where
  \[
  (K)_{ij} = \exp\{- \sum_{l=1}^p \psi_l^2(x_{il} - x_{jl})^2\}
  \]
  and some independent hyper-priors on \( \psi_1, \ldots, \psi_p \) and \( \rho \).
- Log-likelihood function. Marginally, \( y|\sigma^2, \rho, \lambda \sim N(0, \sigma^2(\rho^2 K + I_n)) \) and so
  \[
  p(y|\rho, \lambda) \propto |\rho^2 K + I_n|^{-1/2} \left( 1 + \frac{y^T(\rho^2 K + I_n)^{-1}y}{r\sigma^2_0} \right)^{-\frac{n+1}{2}}
  \]
  and so the log-likelihood function is:
  \[
  \ell(\rho, \lambda) = -\frac{1}{2} \log |\rho^2 K + I_n| - \frac{n + r}{2} \log \left( 1 + \frac{y^T(\rho^2 K + I_n)^{-1}y}{r\sigma^2_0} \right)
  \]
- Low-rank approximation:
  \[
P K P^T = LL^T + D
  \]
  where \( L \) is \( n \times m \) with its first \( m \times m \) block being lower triangular. \( D \) is diagonal with first \( m \) elements zero, the rest are non-negative. \( P \) is a permutation matrix. Some packages, e.g., `inchol()` in `kernlab` returns \( P^T L \) and \( P^T D P \), and no further permutation needs to be done, and one can take \( P = I_n \). In this case, however the resulting \( L \) need not have a lower triangular \( m \times m \) leading block. The low-rank factorization leads to:
  \[
  \rho^2 K + I_n = \rho^2(LL^T + D + \rho^{-2}I_n) = \rho^2(LL^T + A)
  \]
  where \( A = D + \rho^{-2}I_n \).
Sherman-Morrison-Woodbury:
\[
(r^2K + I_n)^{-1} = r^{-2}(LL^T + A)^{-1}
= r^{-2}\{A^{-1} - A^{-1}L(I_m + L^T A^{-1}L)^{-1}L^T A^{-1}\}
= r^{-2}\{A^{-1} - A^{-1}LF^{-1}L^T A^{-1}\}
\]
with \(F = I_m + L^T A^{-1}L\). Also,
\[
\det(r^2K + I_n) = r^{2n} \det(LL^T + A) = r^{2n} \det(A) \det(F)
\]

Low-rank log-likelihood computation: Obtain \(F\) and get its Cholesky factor \(G\), i.e., \(F = GG^T\). Then,
\[
\ell(r, \lambda) = -n \log r - \frac{1}{2} \log \det(A) - \frac{n + r}{2} \log \left\{ 1 + \frac{\tilde{y}^T \tilde{y} - z^T z}{\rho^2 r \sigma_0^2} \right\}
\]
where
\[
\tilde{y} = A^{-1/2}y, \quad z = G^{-1}L^T A^{-1}y
\]

Fitted values. From simple normal theory, \(p(f|y, \sigma^2, r, \lambda) = N(\hat{\mu}_n, \hat{K})\) where,
\[
\hat{\mu}_n = r^2 K (r^2 K + I_n)^{-1} y \\
\hat{K} = \sigma^2\{r^2 K - r^2 K (r^2 K + I_n)^{-1} r^2 K\}
\]
So fitted values equal:
\[
\hat{\mu}_n = K\{A^{-1}y - A^{-1}L(G^T)^{-1}z\}
= (LL^T + D)\{A^{-1}y - A^{-1}L(G^T)^{-1}z\}
= LL^T A^{-1}y + DA^{-1}y - LL^T A^{-1}L(G^T)^{-1}z - DA^{-1}L(G^T)^{-1}z
\]
that is
\[
\hat{\mu}_n = Lw + DA^{-1}y - LF_0 v - DA^{-1}Lv
\]
where
\[
w = L^T A^{-1}y, \quad v = (G^T)^{-1}z, \quad F_0 = L^T A^{-1}L
\]
all of which require \(O(n)\) flops.

For posterior variance, we simplify it to:
\[
\hat{K} = \rho^2 \sigma^2 \{LL^T + D - LL^T (A^{-1} - A^{-1}LF^{-1}L^T A^{-1}) LL^T\}
= \rho^2 \sigma^2 L\{I_m - L^T A^{-1}L + L^T A^{-1}LF^{-1}L^T A^{-1}L\} L^T + \rho^2 \sigma^2 D
= \rho^2 \sigma^2 L\{I_m - F_0 + F_0 F^{-1} F_0\} L^T + \rho^2 \sigma^2 D
\]
• Prediction. Important to append the test cases $x_1^*, \ldots, x_k^*$ etc. to the sequence $(x_1, \ldots, x_n)$ and get a $(n+k) \times m$ incomplete $L$. Can restrict re-ordering to be done only on the first $n$, so that knots are selected from training samples. Let $L_*$ denote the last $k$ rows of this augmented $L$.

$$\hat{\mu}_* = L_* L^T \{ A^{-1} y - A^{-1} L (G^T)^{-1} z \} = L_*(w - F_0 v)$$

$$\hat{K}_* = \rho^2 \sigma^2 L_* \{ I_m - F_0 + F_0 F^{-1} F_0 \} L^T_* + \rho^2 \sigma^2 D_*$$

Hence a random draw from the predictive distribution (given $\sigma^2$ and other covariance parameters) is obtained as: $\hat{\mu}_* + \rho \sigma L_* H_* Z_1 + \rho \sigma Z_2$ where $Z_1$ is drawn from $N_k(0, I_k)$, $Z_2$ is drawn from $N_k(0, D_*)$ and $H_*$ is the left Cholesky factor of $I_k - F_0 + F_0 F^{-1} F_0$.

8 GP regression and uncertainty quantification

It may appear that Gaussian process regression models are primarily motivated by the desire to capture nonlinear relationship between a response and its predictors. However, GP regression, particularly those that use a smooth GP prior, possess a more powerful property in the form of local smoothing. This is apparent in Figure 1, where the width of the prediction band at a test point depends on its proximity to training points. The prediction band is narrower in regions rich in training data, and, wider where training data is scarce. In other words, the GP regression with a smooth (stationary) covariance kernel makes prediction based on local learning (the correlation-range parameter determines localness). Notice that parametric nonlinear models, such as one where throws in second order polynomial terms in a linear regression, cannot perform local learning even though they are capable of capturing non-linearity (to a limited but practically useful extent).

That smooth Gaussian process regression models are able to quantify localized prediction uncertainty comes handy in two very specific applications: computer emulation and causal inference from nonrandomized studies.

8.1 Computer emulation

In computer emulation data is of the form $(x_i, y_i), i = 1, \ldots, n$, where each $x_i \in S \subset \mathbb{R}^d$ is an input to a blackbox computer model (a simulator) and $y_i$ is the corresponding output, which may be a scalar, a vector of a more complex object. Often the simulator is deterministic, that is it produces the same output on repeated runs with the same input point. At other times, the simulator may inject a relatively small random noise. Often times, the simulator runs are expensive and slow and hence it may be desirable to statistically estimate the blackbox input-output relationship from only a small number of training data generated by the simulators. Such a statistical estimate of the simulator is called an emulator.
This is clearly a prediction problem and one that fits nicely within our nonparametric regression framework particularly when the simulator output is univariate\(^{10}\). And it appears a simpler problem since the noise variance is essentially zero. However, computer emulation is distinct from general prediction problems because here one is able to produce new data when and as needed. Indeed, it is natural to ask whether additional training data should be obtained to guarantee a certain estimation accuracy of the emulator. And if additional data is needed, where should one choose the new input points to be?

Addressing the latter question requires localized uncertainty quantification. Naturally GP models for the blackbox function appears extremely appealing\(^{11}\). One can look at input regions where the predictive band is the widest, and, if it is wider than a pre-specified threshold then request news runs of the simulator in these high-uncertainty regions. A more rigorous adaptive, sequential design selection can be performed with the help of a well elicited utility function. For example, if the goal is to estimate the minima of the simulator function \(f\) (up to a given threshold) then one often chooses the next input point as \(x^*\) that maximizes the expected improvement utility function

\[
I(x) = \mathbb{E}\{\max(m - f(x), 0)\}, \\
\text{where } m \text{ is the current estimate of } \min_{x} f(x) \text{ and the} \\
\text{expectation is taken with respect to the posterior distribution of } f \text{ given the current training data (Jones et al., 1998)}.
\]

8.2 Causal inference from nonrandomized studies

Often in non-randomized studies, a strong correlation between a treatment and other predictors\(^{12}\) invalidates standard regression estimates of treatment effects. Such confounding can strongly bias the estimate of the coefficient on the treatment variable, producing unreasonable magnitudes and even the wrong sign.

Regression smoothing offers a strikingly better alternative to quantifying treatment effects in nonrandomized studies. Let \(X = (T, Z)\) where \(T\) is a treatment variable, and \(Z\) is the vector of other, possibly confounding, factors that may influence the response. The population level effect of increasing \(T = t\) to \(T = t + \Delta\) is captured in

\[
\int \{f(t + \Delta, z) - f(t, z)\}dQ^Z(z), \\
\text{and an in-sample version of the quantity is } \psi_n = \frac{1}{n}\sum_{i=1}^{n}\{f(t + \Delta, Z_i) - f(t, Z_i)\}. \\
\text{Inference on } \psi_n \text{ involves predicting the value of } f \text{ at various } (t, Z_i) \text{ combinations. A strong confounding between } T \text{ and } Z \text{ implies that a} \\
\text{certain } (t, Z_i) \text{ pair may be far away from the observed data cloud } (T_i, Z_i)_{i=1}^{n}, \text{ and thus}
\]

\(^{10}\)Even when a simulator produces a vector or function valued output, it often suffices to build emulators for each scalar output component. Unlike in real world data, the simulator output components are deterministically related to each other, and, thus estimating their dependency yields little statistical power while making the estimation problem much harder. See Mengyang Gu’s thesis for a discussion of this on pyroclastic flow simulation.

\(^{11}\)The literature on GP regression based computer emulation is fairly extensive. A very limited sampling is: Sacks et al. (1989), Kennedy and O’Hagan (2001), Bayarri et al. (2009), Gramacy and Lee (2009), Montagna and Tokdar (2016).

\(^{12}\)Here we assume that such confounding predictors are at least measured. Dealing with unmeasured confounders is a much more complex task.
Figure 3: In a toy example a single covariate ‘age’ influences both treatment assignment and a continuous “outcome”; younger people are more likely to receive treatment and higher outcome scores. For either group, counterfactuals are predicted by learning outcome-age relation from the other group’s data and estimated treatment effect (“Effect”), i.e., the difference between predicted outcome and predicted counterfactual outcome, is shown as a function of age. The true effects curve is deliberately omitted to focus on the issues of potential prediction bias arising from confounding. Linear model (LM) fits are good within groups, but appear overconfident while predicting counterfactuals. Add-GP trades potential bias with increased uncertainty bands and produces a more robust effect quantification. BART, which has been used by Hill (2011) for counterfactual prediction, produces shorter error bars and remains prone to bias.

prediction at such a pair is akin to extrapolation, which is always prone to undetectable bias.

Smooth GP regression, where prediction is done via local smoothing, will be sensitive to lack of good data near an isolated \((t, Z_i)\) pair and will respond by increasing the prediction error bar at that point; see Figure 3 for a clear illustration. Such a behavior is conceptually quite similar to the ideas of matching (Rosenbaum and Rubin, 1983; Stuart, 2010), but regression smoothing offers several conceptual and practical advantages: A) it works effortlessly with binary, multi-category, continuous outcomes, and potentially vector valued treatments; B) it requires no formal “balance check”\(^\text{13}\) – which is often hard to implement for high-dimensional \(Z\) and continuous treatment variables (Hill, 2008) – instead lack of local balance is flagged by widening of prediction uncertainty (see Figure 3); C) it does not require removing observation units that

\(^{13}\)For binary treatment \(T \in \{-1, 1\}\) it requires checking that, post-matching, the sample distributions of \(Z|(T = 1)\) and \(Z|(T = -1)\) are close to each other.
contribute to non-overlapping propensity score distributions\textsuperscript{14}, rather such units are viewed as data units that will produce very isolated \((t, Z_i)\) pairs and this incur widest prediction error bars.

References


\textsuperscript{14}Again, for binary treatment, non-overlapping is said to have occurred if the range of \(\{Q(T = 1|Z = Z_i) : T_i = 1\}\) is (substantially) different from the range of \(\{Q(T = 1|Z = Z_i) : T_i = -1\}\).