Chapter 2

Theoretical Development

2.1 Introduction

In this chapter I present a new class of nonstationary correlation functions and determine the smoothness properties of Gaussian processes (GPs) whose correlation functions lie in the class. The new class is a generalization of the kernel convolution covariance of Higdon et al. (1999). Next, I review results on the continuity and differentiability of sample paths from isotropic Gaussian processes based on the characteristics of the correlation functions of the GPs. I apply these results to the generalized kernel convolution correlation functions and show that they retain the smoothness properties of the isotropic correlation functions upon which they are based, provided that the underlying kernel structure is sufficiently smooth. I close by discussing some potential advantages of the generalized kernel convolution correlation.

2.2 Nonstationary Covariance Functions Using Convolutions of Kernels

In this section, I describe in detail the approach of Higdon et al. (1999) (henceforth HSK) for defining nonstationary covariance functions. HSK propose a nonstationary spatial covariance function, $C(\cdot,\cdot)$, defined by

$$C(\boldsymbol{x_i}, \boldsymbol{x_j}) = \int_{\Re^2} K_{\boldsymbol{x_i}}(\boldsymbol{u}) K_{\boldsymbol{x_j}}(\boldsymbol{u}) d\boldsymbol{u}, \tag{2.1}$$

where $\boldsymbol{x_i}$, $\boldsymbol{x_j}$, and $\boldsymbol{u}$ are locations in $\Re^2$, and $K_{\boldsymbol{x}}$ is a kernel function centered at $\boldsymbol{x}$. They motivate this construction as the covariance function of a white noise process, $\psi(\cdot)$, convolved with the kernel function to produce the process, $Z(\cdot)$, defined by

$$Z(\boldsymbol{x}) = \int_{\Re^2} K_{\boldsymbol{x}}(\boldsymbol{u})\psi(\boldsymbol{u})d\boldsymbol{u}.$$

One can avoid the technical details involved in carefully defining such a white noise process by using the definition of positive definiteness to show directly that the covariance function is positive definite in every Euclidean space, $\Re^p, p = 1, 2, \ldots$:

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(\boldsymbol{x_i}, \boldsymbol{x_j}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \int_{\Re^P} K_{\boldsymbol{x_i}}(\boldsymbol{u})K_{\boldsymbol{x_j}}(\boldsymbol{u})d\boldsymbol{u} \\
&= \int_{\Re^P} \sum_{i=1}^{n}\sum_{j=1}^{n} a_i K_{\boldsymbol{x_i}}(\boldsymbol{u}) a_j K_{\boldsymbol{x_j}}(\boldsymbol{u})d\boldsymbol{u} \\
&= \int_{\Re^P} \sum_{i=1}^{n} a_i K_{\boldsymbol{x_i}}(\boldsymbol{u}) \sum_{j=1}^{n} a_j K_{\boldsymbol{x_j}}(\boldsymbol{u})d\boldsymbol{u} \\
&= \int_{\Re^P} \left(\sum_{i=1}^{n} a_i K_{\boldsymbol{x_i}}(\boldsymbol{u})\right)^2 d\boldsymbol{u} \\
&\geq 0. \qquad (2.2)
\end{aligned}$$

The key to achieving positive definiteness is that each kernel is solely a function of its own location. Apart from this restriction, the structure of the kernel is arbitrary. I will return to this proof of positive definiteness when I generalize the HSK approach in Section 2.3.

Next I show the closed form of the HSK covariance for Gaussian kernels based on the equivalence of convolutions of densities with sums of independent random variables:

$$\begin{aligned}
C(\boldsymbol{x_i}, \boldsymbol{x_j}) &= \int K_{\boldsymbol{x_i}}(u) K_{\boldsymbol{x_j}}(\boldsymbol{u})d\boldsymbol{u} \\
&= \int \frac{1}{(2\pi)^{\frac{P}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x_i} - \boldsymbol{u})^T \Sigma_i^{-1}(\boldsymbol{x_i} - \boldsymbol{u})\right) \\
&\quad \times \frac{1}{(2\pi)^{\frac{P}{2}}|\Sigma_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x_j} - \boldsymbol{u})^T \Sigma_j^{-1}(\boldsymbol{x_j} - \boldsymbol{u})\right) d\boldsymbol{u}.
\end{aligned}$$

Recognize the expression as the convolution

$$\int h_{\boldsymbol{A}}(\boldsymbol{u} - \boldsymbol{x_i}) h_{\boldsymbol{U}}(\boldsymbol{u})d\boldsymbol{u} = \int h_{\boldsymbol{A},\boldsymbol{U}}(\boldsymbol{u} - \boldsymbol{x_i}, \boldsymbol{u})d\boldsymbol{u},$$

where $h(\cdot)$ is the normal density function, $\boldsymbol{A} \sim \mathrm{N}(\boldsymbol{0}, \Sigma_i)$, $\boldsymbol{U} \sim \mathrm{N}(\boldsymbol{x_j}, \Sigma_j)$, and $\boldsymbol{A}$ and $\boldsymbol{U}$ are independent. Now consider the transformation $\boldsymbol{W} = \boldsymbol{U} - \boldsymbol{A}$ and $\boldsymbol{V} = \boldsymbol{U}$, which has Jacobian of 1. This gives us the following equalities based on the change of variables:

$$\begin{aligned}\int h_{\boldsymbol{A},\boldsymbol{U}}(\boldsymbol{u} - \boldsymbol{x_i}, \boldsymbol{u})d\boldsymbol{u} &= \int h_{\boldsymbol{W},\boldsymbol{V}}(\boldsymbol{u} - (\boldsymbol{u} - \boldsymbol{x_i}), \boldsymbol{u})d\boldsymbol{u} \\ &= \int h_{\boldsymbol{W},\boldsymbol{V}}(\boldsymbol{x_i}, \boldsymbol{u})d\boldsymbol{u} \\ &= h_{\boldsymbol{W}}(\boldsymbol{x_i}).\end{aligned}$$

Since $\boldsymbol{W} = \boldsymbol{U} - \boldsymbol{A}$, $\boldsymbol{W} \sim \mathrm{N}(\boldsymbol{x_j}, \Sigma_i + \Sigma_j)$ and therefore

$$\begin{aligned}C(\boldsymbol{x_i}, \boldsymbol{x_j}) &= h_{\boldsymbol{W}}(\boldsymbol{x_i}) \\ &= \frac{1}{(2\pi)^{\frac{P}{2}} \mid \Sigma_i + \Sigma_j \mid^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x_i} - \boldsymbol{x_j})^T(\Sigma_i + \Sigma_j)^{-1}(\boldsymbol{x_i} - \boldsymbol{x_j})\right).\end{aligned}$$

Absorbing the necessary constants into the matrices in the quadratic form and dividing by the standard deviation function, $\sigma(\boldsymbol{x_i}) = \frac{1}{2^{\frac{P}{2}}\pi^{\frac{P}{4}}|\Sigma_i|^{\frac{1}{4}}}$, we arrive at the nonstationary correlation function, $R(\cdot,\cdot)$, defined by

$$R(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{2^{\frac{P}{2}}|\Sigma_i|^{\frac{1}{4}}|\Sigma_j|^{\frac{1}{4}}}{|\Sigma_i + \Sigma_j|^{\frac{1}{2}}} \exp\left(-(\boldsymbol{x_i} - \boldsymbol{x_j})^T\left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1}(\boldsymbol{x_i} - \boldsymbol{x_j})\right). \tag{2.3}$$

Examining the exponential and its quadratic form, we see that this is nothing but a squared exponential stationary correlation, but in place of the squared Mahalanobis distance, $\tau^2 = (\boldsymbol{x_i} - \boldsymbol{x_j})^T\Sigma^{-1}(\boldsymbol{x_i} - \boldsymbol{x_j})$, for arbitrary fixed positive definite matrix Σ, we instead use a quadratic form with the average of the kernel matrices for the two locations. If the kernel matrices are constant, we recover the special case of the squared exponential correlation based on Mahalanobis distance. If they are not constant with respect to $\boldsymbol{x}$, the evolution of the kernel covariance matrices in space produces nonstationary covariance. To construct a covariance function, one merely includes a variance function.

Independently, Gibbs (1997, p. 49, equ. 3.82) derived a special case of the HSK covariance function in which the kernel matrices, Σ_i, are taken to be diagonal positive definite matrices. Gibbs (1997) makes an astute observation about the characteristics of the nonstationary covariance model that applies to the HSK covariance and to my generalization of HSK as well (Section 2.3). When the size of the kernels changes quickly, the resulting correlation structure can be counterintuitive because of the function, which Gibbs calls the 'prefactor', in front of the exponential in (2.3). When the kernels centered at $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ are similar in size, the numerator and denominator more or less cancel out, but when one kernel is much larger than the other, the square root of the determinant in the denominator dominates the product of the fourth roots of the determinants in the numerator; this effect causes smaller correlation than achieved based solely on the exponential term. This is most easily seen graphically in a one-dimensional example in Figure 2.1, where I show $R(-0.5, x)$ (the correlation between the point -0.5 and all other points) and $R(0.5, x)$, when the kernel size changes drastically at $x = 0$. We see that the correlation of $x = 0.5$ with the points to its left drops off more quickly than the correlation of $x = -0.5$ with its neighboring points, because of the effect of the prefactor, even though the kernel centered at $x = 0.5$ is large, and the kernel centered at $x = -0.5$ is small. This is counter to intuition and to our goal for the nonstationary function because at certain distances, the correlation between two points whose kernels are relatively small is larger than the correlation between a point whose kernel is small and a point whose kernel is large. For this example, sample functions are least smooth at the x values where the kernel size changes quickly (Figure 2.1d), rather than being least smooth at the x values with the small kernels. This effect seems to be restricted to situations in which the kernel sizes change very quickly, so it may not be material in practice. However, the phenomenon may arise occasionally in sample paths in the regression modelling, as discussed in Section 4.6.1.

2.3 Generalized Kernel Convolution Covariance Functions

One potential drawback to the kernel convolution approach is that the HSK formulation using Gaussian kernels produces a nonstationary covariance with smoothness properties similar to the stationary squared exponential correlation (as shown in Section 2.5.5). In particular, if the kernel matrices vary sufficiently smoothly in the covariate space, then the sample paths based on the nonstationary covariance are infinitely differentiable. Stein (1999) discusses in detail why such highly smooth paths are undesirable and presents an asymptotic argument for using covariance functions in which the smoothness is allowed to vary. In hopes of avoiding such a high degree of smoothness, one might think of extending the HSK approach by using non-Gaussian kernels, but unless the convolution (2.1) can be done in closed form, this would entail numerical integration.

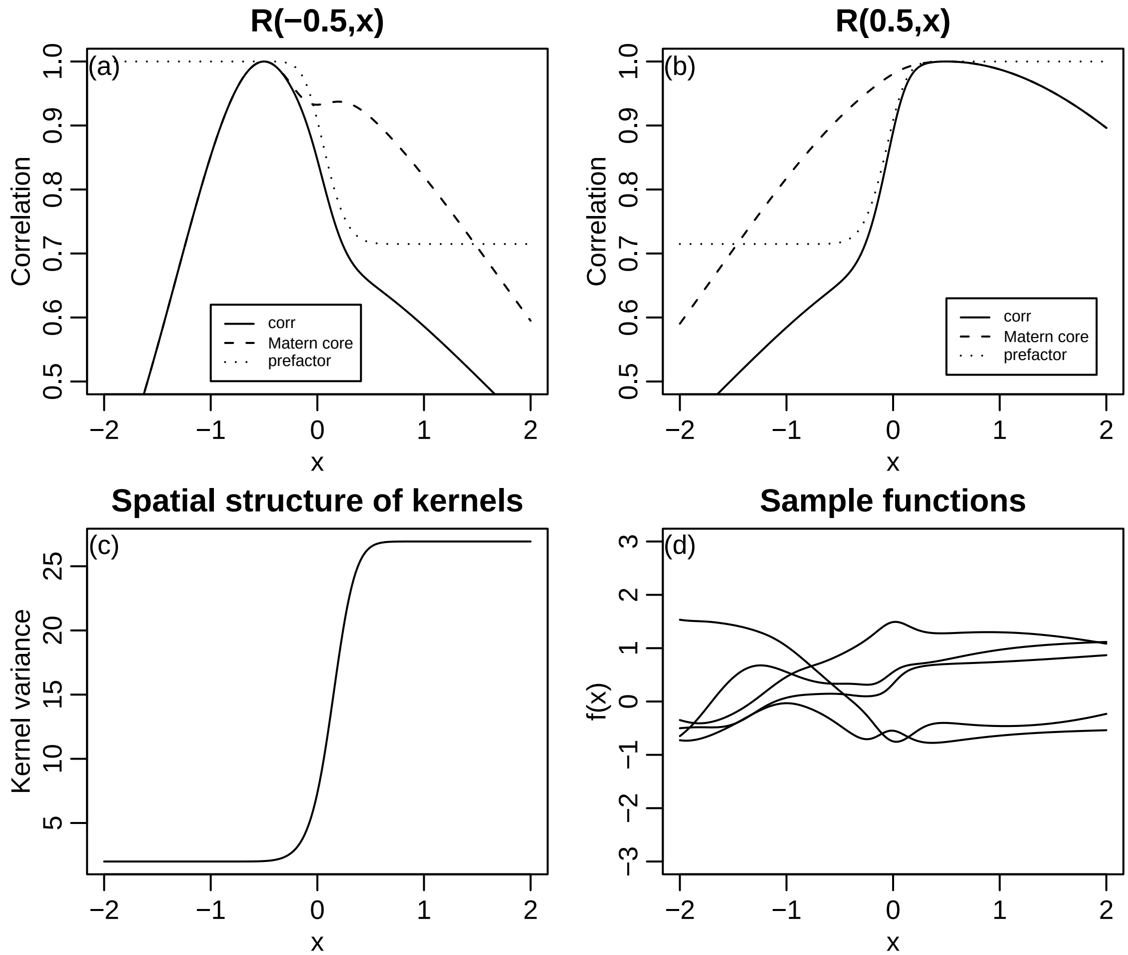

Figure 2.1. (a) Correlation of $f(-0.5)$ with the function at all other points. (b) Correlation of $f(0.5)$ with the function at all other points. (c) Kernel variance as a function of x. (d) Five sample functions drawn from the Gaussian process distribution; notice that the functions are least smooth at the location of the sharp change in the kernel size.

In this section I extend the HSK covariance in a way that provides a closed form correlation function. I produce a class of nonstationary correlation functions that provide more flexibility than the HSK formulation. Consider the quadratic form,

$$Q_{ij} = (\boldsymbol{x_i} - \boldsymbol{x_j})^T \left(\frac{\Sigma_i + \Sigma_j}{2} \right)^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j}), \tag{2.4}$$

at the heart of the correlation function (2.3) constructed via the kernel convolution. We have seen that the HSK nonstationary correlation function is nothing but the squared exponential correlation with this new quadratic form in place of a Mahalanobis distance. This relationship raises the possibility of producing a nonstationary version of an isotropic correlation function by using Q_{ij} in place of $\tau^2 = \boldsymbol{\tau}^T\boldsymbol{\tau}$ in the isotropic function. In practice, one uses $\frac{\tau}{\kappa} = \sqrt{Q_{ij}}$, since the scale parameter κ is redundant and can be absorbed into the kernel matrices, Σ_i and Σ_j, which are allowed to vary in size during the modelling anyway. The following general result applies in particular to correlation functions that are positive definite in Euclidean space of every dimension, in particular the power exponential, rational quadratic, and Matérn correlation functions (1.2-1.4).

Theorem 1 *If an isotropic correlation function, $R(\tau)$, is positive definite on $\Re^p$ for every $p = 1, 2, \ldots$, then the function, $R(\cdot,\cdot)$, defined by*

$$R(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{2^{\frac{p}{2}} |\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}}}{|\Sigma_i + \Sigma_j|^{\frac{1}{2}}} R\left(\sqrt{Q_{ij}}\right) \tag{2.5}$$

with $\sqrt{Q_{ij}}$ used in place of τ, is positive definite on $\Re^p$, $p = 1, 2, \ldots$, and is a nonstationary correlation function.

Proof: The proof is a simple application of Theorem 2 of Schoenberg (1938, p. 817), which states that the class of functions positive definite on Hilbert space is identical with the class of functions of the form,

$$R(\tau) = \int_0^\infty \exp\left(-\tau^2 s\right) dH(s), \tag{2.6}$$

where $H(\cdot)$ is non-decreasing and bounded and $s \geq 0$. The class of functions positive definite on Hilbert space is identical to the class of functions that are positive definite on $\Re^p$ for $p = 1, 2, \ldots$ (Schoenberg 1938). We see that the covariance functions in this class are scale mixtures of the squared exponential correlation. The underlying stationary correlation function with argument $\sqrt{Q_{ij}}$ can be expressed as

$$\begin{aligned} R\left(\sqrt{Q_{ij}}\right) &= \int_0^\infty \exp\left(-Q_{ij}s\right) dH(s) \\ &= \int_0^\infty \exp\left(-(\boldsymbol{x_i} - \boldsymbol{x_j})^T \left(\frac{\frac{\Sigma_i}{s} + \frac{\Sigma_j}{s}}{2}\right)^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j})\right) dH(s) \\ &= \int_0^\infty \int_{\Re^P} K_{\boldsymbol{x_i},s}(\boldsymbol{u}) K_{\boldsymbol{x_j},s}(\boldsymbol{u}) d\boldsymbol{u} dH(s). \end{aligned}$$

Since s is non-negative, it becomes part of the kernel matrices, and the last expression can be seen to be positive definite based on (2.2).
Q.E.D.

This approach replaces the kernel at each location with a scale mixture of kernels where a common scale is used for all the locations (See Matérn (1986, pp. 32-33) for some discussion of generating new stationary correlation functions as scale mixtures of stationary correlation functions.) Using different distributions, H, for the scale parameter, S, produces different nonstationary correlation functions. A nonstationary version of the rational quadratic correlation function of the form (2.5) is

$$R(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{2^{\frac{P}{2}} |\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}}}{|\Sigma_i + \Sigma_j|^{\frac{1}{2}}} \left(\frac{1}{1+Q_{ij}}\right)^{\nu}.$$

This can be seen to be of the scale mixture form by taking $S \sim \Gamma(\nu, 1)$,

$$\int \exp(-Q_{ij}s)dH(s) = E(\exp(-Q_{ij}s)) = M_S(-Q_{ij}; \nu, 1) = \left(\frac{1}{1+Q_{ij}}\right)^{\nu},$$

where M_S is the moment generating function of S. This makes sense since the rational quadratic correlation function has the form of a t density, which is a mixture of Gaussians with an inverse gamma distribution for the variance of the Gaussian, which is proportional to $\frac{1}{S}$. A nonstationary version of the Matérn correlation function is

$$R(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{2^{\frac{P}{2}} |\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}}}{|\Sigma_i + \Sigma_j|^{\frac{1}{2}}} \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\sqrt{2\nu Q_{ij}}\right)^{\nu} K_{\nu}\left(\sqrt{2\nu Q_{ij}}\right). \tag{2.7}$$

Using an integral expression for the Bessel function (Gradshteyn and Ryzhik 1980, p. 340, equ. 9; McLeish 1982), one can easily show that in this case S is distributed inverse-gamma $(\nu, 1/4)$. In Section 2.5.4 (stationary) and Section 2.5.5 (nonstationary), I show that the existence of moments of S is directly related to the existence of mean square and sample path derivatives of processes whose covariance is produced by mixing a squared exponential covariance over the scale parameter. Rather than producing a closed form nonstationary correlation function by substituting the quadratic form (2.4) into an isotropic correlation function, one can instead construct nonstationary correlation functions (possibly without closed form) by choosing a distribution over the scale parameter, with the distribution chosen to produce the desired smoothness properties based on the moments of the distribution.

The Matérn correlation function is proportional to the Bessel density function (McLeish 1982). Based on the fact that convolutions of Bessel densities are also Bessel densities (McLeish 1982; Matérn 1986, pp. 29-30), we might expect that the Matérn nonstationary correlation (2.7) could be derived directly from a convolution of Bessel densities. However, the Bessel distribution is closed under convolution only for fixed scale parameters (this can be seen by multiplying two t densities, since the characteristic function of the Bessel density is proportional to a t density), so it does not directly correspond to a convolution of the type (2.1) for which $\Sigma_i \neq \Sigma_j$.

The quadratic form (2.4) defines a semi-metric space (Schoenberg 1938), in which the distance function is $(\boldsymbol{x_i'} - \boldsymbol{x_j'})^T(\boldsymbol{x_i'} - \boldsymbol{x_j'})$, where $\boldsymbol{x_i'} = \left(\frac{\Sigma_i+\Sigma_j}{2}\right)^{-\frac{1}{2}} \cdot \boldsymbol{x_i}$. However, the new location, $\boldsymbol{x_i'}$, varies depending on the other point, $\boldsymbol{x_j}$, through its dependence on Σ_j. The distance function violates the triangle inequality, even if one considers the points as lying in a higher dimensional space, so the space is not an inner-product space. To see this, consider a one-dimensional example with three points on a line, two points equidistant from the central point and on either side, $x_1 = -1, x_2 = 0, x_3 = 1$. Let the Gaussian kernel at the center point decay slowly along the line, $K_{x_2}(x) = \phi(x; -1, 3^2)$ while the two other Gaussian kernels decay more quickly along the line, $K_{x_1}(x) = \phi(x; -1, 1), K_{x_3}(x) = \phi(x; 1, 1)$. The distance between the central point and either side point is then 0.2, which smaller than half the distance, 4, between the two side points.

To construct the nonstationary correlation functions introduced here, we need kernels at all locations in the space $\mathcal{X}$. As described in Section 3.2, the kernels are modelled as functions of stochastic processes that determine the kernel eigenvectors and eigenvalues. This induces stochastic processes for the elements of the kernel matrices. As I will discuss in Section 2.5.5, the smoothness properties of these elements in part determine the smoothness of stochastic processes parameterized by the nonstationary correlation introduced here.

2.4 Nonstationary Covariance on the Sphere

The generalized kernel convolution covariance model can be extended for use on the sphere, S^2, and other non-Euclidean spaces. On the sphere, the equivalence of translation and rotation causes difficulty in defining kernels that produce correlation behavior varying with direction. The following recipe allows one to create a nonstationary model for the sphere. First, define a truncated

Gaussian kernel at each location in a Euclidean projection of the sphere centered at that location. Let the value of the kernel be zero for distances at which the Euclidean approximation to angular distance is poor. As usual, let the kernels vary smoothly in space. Project the kernels from the Euclidean projection back to the spherical domain, thereby defining a set of kernels on S^2, one kernel at each location of interest. Then define the correlation function as the convolution in the spherical domain. The key to showing positive definiteness (2.2) of the kernel convolution covariance is that each kernel is solely a function of the location of the kernel; this approach satisfies that condition. The additional integration over a scale parameter can also be done here in the spherical domain to produce a class of nonstationary correlation functions on the sphere. In practice, as described in Section 5.4.1.4, I have calculated the correlations directly in the Euclidean projections with untruncated kernels so as to be able to use the analytic form for the correlation (2.5). I found that the resulting correlation, although not guaranteed to be positive definite, does not cause numerical problems. Note that smoothness properties of processes on S^2 follow from those of processes on $\Re^3 \supset S^2$ if one chooses a covariance function that is positive definite on $\Re^3$ and sets $\tau_{ij} = 2\sin\left(\frac{\rho_{ij}}{2}\right)$, where ρ_{ij} is the angular distance between locations $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$.

There has been little work on nonstationary covariance modelling on the sphere apart from a nonlinear mapping approach (Das 2000) that extended the work of Sampson and Guttorp (1992). Das (2000) mapped the original sphere to a new sphere in which stationarity is assumed to hold and then used stationary covariance models valid on S^2.

2.5 Smoothness Properties of Covariance Functions

2.5.1 Overview

The functional form and parameter values of the covariance function of a Gaussian process distribution determine the smoothness properties of the process and sample paths drawn from the distribution. Covariance functions can give Gaussian processes whose sample paths range from discontinuous to analytic. While data can inform the choice of covariance function to some degree, this decision is also a philosophical choice based on one's conception of the underlying physical or scientific process.

Two important characteristics of stochastic processes are mean square properties and sample path properties, which I define in Section 2.5.2. All of the Gaussian processes I consider here are both mean square and sample path continuous based on simplifications of the arguments given here. The key difference amongst correlation functions lies in the differentiability properties associated with them. Using results from the stochastic process literature, in Section 2.5.4 I derive the mean square and sample path differentiability properties of stochastic processes parameterized by the stationary correlation functions (1.2-1.4). These results are well-known but are not collected or proven in one place to my knowledge. In particular, sample path properties of familiar isotropic correlation functions are relatively little discussed (but see Abrahamsen (1997)). I present the material here because the smoothness properties associated with the nonstationary kernel convolution correlation functions (Section 2.5.5) follow from those of the underlying isotropic correlation functions on which they are based. I focus particularly on sample path properties, because I believe these are most relevant when selecting a correlation function. The analyst is more likely to be able to make some intuitive judgement about sample path properties of the process at hand than about mean square properties. Mean square properties are easier to derive, being directly related to derivatives of the covariance function and moments of the spectral distribution, and much of the literature concentrates on these (e.g., Stein (1999)). Even if one is not directly interested in mean square properties, they are useful as a first step in determining sample path properties, as we will see. At times hereafter I refer to mean square and sample path properties of a correlation function, by which I mean properties of mean-zero stochastic processes with the given correlation function. For sample path properties, the results hold only for Gaussian processes.

2.5.2 Theoretical framework

First I give brief definitions of mean square and sample path continuity and differentiability, following Adler (1981). In the remaining sections of this chapter, I will use $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{u}$ to indicate locations in the covariate space and x_p to indicate the pth scalar element of $\boldsymbol{x}$. In the stationary case, let $\boldsymbol{\tau} = \|\boldsymbol{x} - \boldsymbol{y}\|$, and in the isotropic case, let $\tau = \|\boldsymbol{\tau}\| = \sqrt{\boldsymbol{\tau}^T\boldsymbol{\tau}}$. This avoids having double subscripts indicating both location and covariate. However as a notational exception for the following definition, let $\boldsymbol{x_i}, i = 1, 2, \ldots$ be a sequence of locations such that $\|\boldsymbol{x_i} - \boldsymbol{x}\| \to 0$

as $i \to \infty$. If $Z(\boldsymbol{x_i}) \overset{m.s.}{\to} Z(\boldsymbol{x})$ as $i \to \infty$, then $Z(\cdot)$ is continuous in mean square at $\boldsymbol{x}$. If the convergence is almost sure, then $Z(\cdot)$ is almost surely continuous at $\boldsymbol{x}$. If there exists $Z_p^{(1)}(\boldsymbol{x})$ such that

$$\frac{Z(\boldsymbol{x}+\epsilon \boldsymbol{u_p}) - Z(\boldsymbol{x})}{\epsilon} \overset{m.s.}{\to} Z_p^{(1)}(\boldsymbol{x}), \epsilon \to 0$$

where $\boldsymbol{u_p}$ is the unit vector in the pth direction, then $Z_p^{(1)}(\boldsymbol{x})$ is the mean square derivative of $Z(\cdot)$ at $\boldsymbol{x}$. Again if the convergence is almost sure, then $Z_p^{(1)}(\boldsymbol{x})$ is the pth-order almost sure partial derivative at $\boldsymbol{x}$. Mth-order partial derivatives, either mean square or almost sure, can be defined in similar fashion. If almost sure continuity or differentiability hold simultaneously with probability one for all $\boldsymbol{x} \in I \subset \Re^P$ then $Z(\cdot)$ is sample path continuous or has a sample path partial derivative, respectively, on I. Sample path differentiability involves the existence and continuity of sample path partial derivatives, as I will discuss shortly.

In general, because the finite dimensional distributions of a stochastic process do not determine the sample path properties of the process, showing sample path continuity or differentiability relies on the notion of separability pioneered by Doob (1953, pp. 51-53) and discussed in detail in Gihman and Skorohod (1974, p. 164) and Adler (1981, p. 14). A separable process is one for which the finite-dimensional distributions determine the sample path properties. By virtue of Theorem 2.4 of Doob (1953, p. 57), for any stochastic process, $\tilde{Z}(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}$ with $\mathcal{X}$ a linear space, there exists a version of the process, $Z(\boldsymbol{x})$, that is separable and is stochastically equivalent to the original process. Once one assumes that one is working with the separable version of the stochastic process, almost sure continuity (differentiability) can be extended to sample path continuity (differentiability) because the probability one statement at individual points holds simultaneously on a dense countable set of points, and the sample path properties of the separable process are determined by the properties of the process on the dense countable set. From this point forward, I will assume all processes are separable.

Mean square properties of correlation functions are frequently analyzed, in part because they can be readily determined from the correlation function, or for stationary correlation functions, from the spectral representation of the correlation function. A process is mean square continuous at $\boldsymbol{u}$ if and only if the covariance function $C(\boldsymbol{x}, \boldsymbol{y})$ is continuous at $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{u}$ (Cramér and Leadbetter 1967, p. 83; Loève 1978, p. 136; Adler 1981, p. 26). All the correlation functions considered in this work are continuous, and therefore processes with these correlation functions are mean square continuous. Mean square differentiability is directly related to the existence of derivatives of the correlation function. A random field, $Z(\cdot)$, has mean square partial derivative at $\boldsymbol{u}$, $Z_p^{(1)}(\boldsymbol{u})$ if and only if $\partial^2 C(\boldsymbol{x},\boldsymbol{y})/(\partial x_p \partial y_p)$ exists and is finite at $(\boldsymbol{u},\boldsymbol{u})$ (Adler 1981, p. 27). If the derivative exists, the covariance function of the mean square partial derivative process is the partial derivative of the original covariance function. For stationary processes, one need only consider the partial derivative evaluated at $\mathbf{0}$, and the correlation function of the partial derivative process is $\partial^2 C(\boldsymbol{\tau})/\partial \tau_p^2$. The existence of Mth-order mean square partial derivatives is equivalent to the finiteness of the relevant $2M$th-order partial derivatives of the covariance function (Adler 1981, p. 27; Vanmarcke 1983, p. 111),

$$\frac{\partial^{2M} C(\boldsymbol{x},\boldsymbol{y})}{\partial x_{p_1}\cdots\partial x_{p_M}\partial y_{p_1}\cdots\partial y_{p_M}},$$

evaluated at $(\boldsymbol{u},\boldsymbol{u})$ for $p_m \in \{1,\ldots,P\}, m \in \{1,\ldots,M\}$.

Stationary covariance functions can be expressed as the Fourier transform of the spectral distribution, $H(\cdot)$,

$$C(\boldsymbol{\tau}) = \int_{\Re^P} \exp(i\boldsymbol{w}^T\boldsymbol{\tau})dH(\boldsymbol{w}).$$

If $H(\cdot)$ is absolutely continuous with respect to Lebesgue measure, then the spectral density, $h(\cdot)$, exists and can be expressed as

$$h(\boldsymbol{w}) = \frac{1}{(2\pi)^P}\int_{\Re^P} \exp(-i\boldsymbol{w}^T\boldsymbol{\tau})C(\boldsymbol{\tau})d\boldsymbol{\tau}.$$

In other words, the covariance function is the characteristic function of the distribution, $H(\cdot)$. In the isotropic case, $H(\boldsymbol{w})$ depends only on $\|\boldsymbol{w}\|$ (Adler 1981, p. 35). Using the well-known relationship between derivatives of a characteristic function evaluated at the origin and moments of the distribution, one can see that mean square differentiability is equivalent to the existence of moments of the spectral distribution. A process on $\Re^1$ is mean square differentiable if and only if

$$\int w^2 dF(w) < \infty,$$

because the existence of the second moment is equivalent to having two derivatives of the covariance at the origin (Stein 1999, p. 27). As mentioned above, in higher dimensions, the existence of a $2M$th-order partial derivative of the covariance at the origin is equivalent to having the respective Mth-order mean square partial derivative exist. This is also equivalent to the existence of the $2M$th-order spectral moments (Adler 1981, p. 31),

$$(-1)^M \left. \frac{\partial^{2M} C(\boldsymbol{\tau})}{\partial \tau_{p_1}^2 \cdots \partial_{p_M}^2} \right|_{\boldsymbol{\tau}=\mathbf{0}} = \int_{\Re^P} w_{p_1}^2 \cdots w_{p_M}^2 dH(\boldsymbol{w}) < \infty, \tag{2.8}$$

for $p_m \in \{1, \ldots, P\}, m \in \{1, \ldots, M\}$. While the spectral relationship is very useful for assessing mean square differentiability in the stationary case, it does not provide the covariance function of the mean square derivative, only the value of the covariance at $\mathbf{0}$.

Mean square properties are important in part because they are useful for showing sample path properties. In the discussion that follows in this paragraph, except where noted, the cited results are for processes on $\Re^1$. While I have not seen the results shown formally on $\Re^P$, I presume they hold there as well. Cambanis (1973, Theorem 6) has shown that a real, separable, measurable Gaussian process that is not mean square differentiable at any point has with probability one paths that are almost nowhere differentiable. Since mean square differentiability is generally straightforward to determine, the more difficult cases involve showing sample path differentiability for processes that are mean square differentiable. Doob (1953, p. 536) shows that for a separable process, sample functions of the process are absolutely continuous, and hence the functions have derivatives almost everywhere. Furthermore, the mean square derivative process is equal to the sample path derivative process with probability one (Doob 1953, p. 536; Cramér and Leadbetter 1967, p. 85; Yaglom 1987, p. 67). In the Gaussian case, on $\Re^P$, the derivative processes are also Gaussian processes, and the joint distributions of all of these processes are Gaussian (Adler 1981, p. 32). For a function of P variables, if all first-order partial derivatives, $Z_p^{(1)}(\cdot), p = 1, \ldots, P$, exist and are continuous, the function is continuously differentiable and this is sufficient for the function to be first-order differentiable (Olmsted 1961, p. 267; Leithold 1968, pp. 795-796). Since the partial derivatives are themselves functions on $\Re^P$, higher-order derivatives are defined recursively by differentiating the lower-order derivative functions. I demonstrate Mth-order sample path differentiability by showing that all Mth-order partial derivative processes, $Z_{p_1 \cdots p_M}^{(M)}(\cdot)$ for $p_m \in \{1, \ldots, P\}, m \in \{1, \ldots, M\}$ exist and are sample path continuous.

Sample path continuity is difficult to demonstrate in the non-Gaussian case. Adler (1981, p.

48) gives the condition for a stationary process that for $\alpha > 0$ and $\epsilon > \alpha$, if

$$E|Z(\boldsymbol{x}+\boldsymbol{\tau}) - Z(\boldsymbol{x})|^{\alpha} \leq \frac{c\|\boldsymbol{\tau}\|^{2P}}{|\log \|\boldsymbol{\tau}\||^{1+\epsilon}},$$

then $Z(\cdot)$ will be sample path continuous over any compact set in $\Re^P$. In the Gaussian case, the conditions are less strict. If $Z(\cdot)$ is a zero-mean Gaussian process with continuous covariance and for some finite $c > 0$ and some $\epsilon > 0$,

$$E|Z(\boldsymbol{x}) - Z(\boldsymbol{y})|^2 \leq \frac{c}{|\log \|\boldsymbol{x}-\boldsymbol{y}\||^{1+\epsilon}} \tag{2.9}$$

for $\boldsymbol{x}$ and $\boldsymbol{y}$ in I, then the process has continuous sample paths on I (Adler 1981, p. 60). In the stationary case, the condition simplifies to

$$C(\mathbf{0}) - C(\boldsymbol{\tau}) \leq \frac{c}{|\log \|\boldsymbol{\tau}\||^{1+\epsilon}}. \tag{2.10}$$

Adler (1981, p. 64) also gives a condition for sample path continuity for stationary, zero-mean Gaussian processes based on the spectral representation of the covariance. If for some $\epsilon > 0$,

$$\int_{\Re^P} |\log(1 + \|\boldsymbol{w}\|)|^{1+\epsilon} dH(\boldsymbol{w}) < \infty,$$

then the process is sample path continuous. However, I have not been able to use this spectral condition to demonstrate the continuity of derivative processes.

To summarize, the steps involved in proving sample path differentiability for Gaussian processes are as follows. I focus only on correlation functions in the sections that follow, assuming that the variance function is either constant or has sample paths as smooth as those based on the correlation function. First I show that Mth-order mean square differentiability holds (using either the derivative of the correlation or moments of the spectral distribution) and determine the covariance of the Mth-order mean square partial derivative processes (using the derivatives of the correlation). These mean square derivative processes are probabilistically equivalent to the sample path derivative processes. Then I show that all the Mth-order derivative processes are sample path continuous based on their covariance functions and either condition (2.9) or (2.10).

Processes that are infinitely mean square differentiable may also be mean square analytic. Loève (1978, p. 137) and Stein (1999, p. 33) state that processes on $\Re^1$ are mean square analytic if the covariance function $C(x, y)$ is analytic at (u, u). Both the squared exponential and rational quadratic correlation functions (despite its having a second parameter) are analytic, as I show in Section 2.5.4.4 by demonstrating complex differentiability of the correlation function. This suggests that the sample paths for processes with these correlation functions will be analytic, although I have not seen this result proven. Stein (1999) argues that analytic sample paths are unrealistic for physical processes, since an analytic function is fully determined by its values in a small interval.

2.5.3 Lemmas for proofs

Before venturing into the details of the smoothness properties of Gaussian processes based on correlation function properties, I provide definitions and lemmas that I will use in the proofs. I state the lemmas here for $x_p - y_p$ and $\|\boldsymbol{x}-\boldsymbol{y}\|$ but in the stationary setting, I use $\tau_p = x_p - y_p$ and $\|\boldsymbol{\tau}\| = \|\boldsymbol{x}-\boldsymbol{y}\|$.

First I introduce some notation involving partial derivatives, which will clarify the arguments involved in the remainder of this chapter. The classes $\mathcal{D}^{(m)}(\cdot)$ are defined recursively as follows. Let

$$\mathcal{D}^{(1)}(\cdot) = \left\{ \frac{\partial(\cdot)}{\partial x_1}, \ldots, \frac{\partial(\cdot)}{\partial x_P}, \frac{\partial(\cdot)}{\partial y_1}, \ldots, \frac{\partial(\cdot)}{\partial y_P} \right\}.$$

I will use $D^{(1)}(\cdot) \in \mathcal{D}^{(1)}(\cdot)$ to represent a term taking the form of an element in the class. Then for $m > 1$ let $\mathcal{D}^{(m)}(\cdot) = \left\{ D^{(1)}(D^{(m-1)}(\cdot)) \right\}$. For $m = 2$ we have

$$\mathcal{D}^{(2)}(\cdot) = \left\{ \frac{\partial^2(\cdot)}{\partial x_1^2}, \frac{\partial^2(\cdot)}{\partial x_1 \partial x_2}, \ldots, \frac{\partial^2(\cdot)}{\partial x_1 \partial y_P}, \ldots, \frac{\partial^2(\cdot)}{\partial y_P^2} \right\}.$$

To denote partial derivatives with respect to $\boldsymbol{x}$ only, I use $D_{\boldsymbol{x}}^{(m)}$ and for particular partial derivatives with respect to a coordinate of $\boldsymbol{x}$ I use $D_{x_i}^{(m)}$.

Definition 2 *I define $g(\boldsymbol{x},\boldsymbol{y})$ to be $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^a)$, denoted $g(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^a)$, if there exists $c > 0$ such that for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in a region I,*

$$\frac{g(\boldsymbol{x},\boldsymbol{y})}{\|\boldsymbol{x}-\boldsymbol{y}\|^a} \leq c.$$

Schervish (1995, p. 394) provides an alternate definition, with more extensive notation and related properties, based on a sequence of numbers, r_n, replacing $\|\boldsymbol{x}-\boldsymbol{y}\|$ and the region I. Some properties that will be useful are

- *If* $g_1(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_1})$ *and* $g_2(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_2})$ *then* $g_1(\boldsymbol{x},\boldsymbol{y}) \cdot g_2(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_1+a_2})$.
- *If* $g(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_1})$ *then* $g(\boldsymbol{x},\boldsymbol{y})^{a_2} = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_1 a_2})$.
- *If* $g_1(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_1})$ *and* $g_2(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_2})$*, with* $a_1 \le a_2$*, then* $g_1(\boldsymbol{x},\boldsymbol{y}) + g_2(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^{a_1})$.

Next I prove a series of lemmas.

Lemma 3 $x_p - y_p = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$

Proof: First, the claim is equivalent to $(x_p - y_p)^2 = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^2)$ by the properties of $O_I(\cdot)$. The following bound holds:

$$\frac{(x_p - y_p)^2}{\|\boldsymbol{x}-\boldsymbol{y}\|^2} = \frac{(x_p - y_p)^2}{\sum_q (x_q - y_q)^2} \le 1.$$

Q.E.D.

Lemma 4 *If* $D^{(1)}(g(\boldsymbol{x},\boldsymbol{y}))$ *exists, then* $g(\boldsymbol{x},\boldsymbol{x}) - g(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$.

Proof: Consider

$$\frac{g(\boldsymbol{x},\boldsymbol{y}) - g(\boldsymbol{x},\boldsymbol{x})}{\|\boldsymbol{x}-\boldsymbol{y}\|}. \tag{2.11}$$

By standard results in advanced calculus texts, such as Buck (1965, p. 243) or Leithold (1968, p. 795), if the function $g(\cdot)$ is a differentiable function of $\boldsymbol{y}$, then we can express the numerator as

$$g(\boldsymbol{x},\boldsymbol{y}) - g(\boldsymbol{x},\boldsymbol{x}) = D^{(1)}_{y_1}(g(x,y))(x_1 - y_1) + \cdots + D^{(1)}_{y_P}(g(x,y))(x_P - y_P) + R(x_1 - y_1, \ldots, x_P - y_P),$$

where

$$\lim_{\|\boldsymbol{x}-\boldsymbol{y}\| \to 0} \frac{R(x_1 - y_1, \ldots, x_P - y_P)}{\|\boldsymbol{x}-\boldsymbol{y}\|} = 0.$$

Therefore the expression (2.11) can be bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ in a region I based on the above limit and on lemma 3.

Q.E.D.

Lemma 5 *If* $g(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^a)$ *for* $a > 0$*, then* $g(\boldsymbol{x},\boldsymbol{y}) |\log \|\boldsymbol{x}-\boldsymbol{y}\||^{1+\epsilon}$ *is bounded for* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *in a region* I.

Proof:

$$
\begin{aligned}
g(\boldsymbol{x},\boldsymbol{y})|\log\|\boldsymbol{x}-\boldsymbol{y}\||^{1+\epsilon} &= \frac{g(\boldsymbol{x},\boldsymbol{y})}{\|\boldsymbol{x}-\boldsymbol{y}\|^a}\|\boldsymbol{x}-\boldsymbol{y}\|^a|\log\|\boldsymbol{x}-\boldsymbol{y}\||^{1+\epsilon} \\
&= \frac{g(\boldsymbol{x},\boldsymbol{y})}{\|\boldsymbol{x}-\boldsymbol{y}\|^a}\left|\|\boldsymbol{x}-\boldsymbol{y}\|^{\frac{a}{1+\epsilon}}|\log\|\boldsymbol{x}-\boldsymbol{y}\||\right|^{1+\epsilon}.
\end{aligned}
$$

The fraction is bounded by the assumption that $g(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^a)$. The term inside the outer absolute value has limit of 0 as $\|\boldsymbol{x}-\boldsymbol{y}\| \rightarrow 0$ by Abramowitz and Stegun (1965, p. 68,equ. 4.1.31) and hence is bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ in a region I, and the constant power, $1+\epsilon$, does not affect the boundedness.

Q.E.D.

Lemma 6 *If the kernel matrix elements are continuous,* $Q_{xy} = (\boldsymbol{x}-\boldsymbol{y})^T\left(\frac{\Sigma_x+\Sigma_y}{2}\right)^{-1}(\boldsymbol{x}-\boldsymbol{y})$ *is* $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^2)$. *If the kernel matrix elements are once sample path differentiable, then* $D^{(1)}(Q_{xy})$ *is* $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$. *If the kernel matrix elements are* m *times sample path differentiable, then* $D^{(m)}(Q_{xy}) = O_I(1)$.

Proof: First consider Q_{xy} and absorb the divisor of 2 in the matrix inverse into the kernel matrices. By continuity of the kernel matrix elements, the elements of $\Sigma_x+\Sigma_y$ are bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ in a region I. Expressing $Q_{xy} = \sum_i\sum_j(x_i-y_i)(x_j-y_j)(\Sigma_x+\Sigma_y)^{-1}_{ij}$, it is clear that $Q_{xy} = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^2)$. Next, without loss of generality, consider the first partial derivative $D^{(1)}_{x_i}(Q_{xy}) = c_1\sum_p(x_p-y_p)(\Sigma_x+\Sigma_y)^{-1}_{ip} + (\boldsymbol{x}-\boldsymbol{y})^T D^{(1)}_{x_i}(\Sigma_x+\Sigma_y)^{-1}(\boldsymbol{x}-\boldsymbol{y})$. The first term is $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$, assuming the kernel matrix elements are continuous. For the second term,

$$
\frac{\partial(\Sigma_x+\Sigma_y)^{-1}}{\partial x_i} = (\Sigma_x+\Sigma_y)^{-1}\frac{\partial(\Sigma_x+\Sigma_y)}{\partial x_i}(\Sigma_x+\Sigma_y)^{-1},
$$

so the second term is $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|^2)$ provided the kernel matrix elements are once differentiable. Hence the sum is $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$. Finally consider

$$
D^{(m)}(Q_{\boldsymbol{xy}}) = c_1 D^{(m-2)}\left((\Sigma_x+\Sigma_y)^{-1}\right) + \ldots + c_2(\boldsymbol{x}-\boldsymbol{y})^T D^{(m)}\left((\Sigma_x+\Sigma_y)^{-1}\right)(\boldsymbol{x}-\boldsymbol{y}).
$$

All the terms are bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ in a region I if the kernel matrix elements are at least m times sample path differentiable, so the sum is $O_I(1)$.

Q.E.D.

Lemma 7 *If $ES^{2M} < \infty$, one can interchange differentiation and integration of*

$$\frac{\partial}{\partial Q} \int S^{2M-m} \exp(-Qs) dH(s)$$

for $m \in \{1, \ldots, 2M\}$.

Proof: Corollary 2.4.1 of Casella and Berger (1990, p. 71) states that to interchange differentiation and integration, it is sufficient that for some $\epsilon > 0$, there exists a function $g_0(Q, s)$, integrable with respect to s, such that

$$\left| \frac{\partial g(Q,s)}{\partial Q} |_{Q=Q_0} \right| \leq g_0(Q,s), \forall Q_0 \; s.t. \; |Q_0 - Q| \leq \epsilon$$

where the dominating function $g_0(Q, s)$ is integrable with respect to s. In this case, we have

$$\begin{aligned} \left| \frac{\partial}{\partial Q} g(Q,s)|_{Q=Q_0} \right| &= \left| \frac{\partial}{\partial Q} s^{2M-m} \exp(-Qs)|_{Q=Q_0} \right| \\ &= s^{2M-m+1} \exp(-Q_0 s) \\ &\leq s^{2M-m+1} \\ &= g_0(Q,s), \end{aligned}$$

where the function g_0 does not involve Q and hence is integrable for all Q for $m \in \{1, \ldots, 2M\}$ with respect to $H(s)$ by the assumption that $ES^{2M} < \infty$. Note that differentiating with respect to a function, $a(Q)$ merely introduces the multiplicative factor $\partial Q / \partial a(Q)$, which is constant with respect to s.

Q.E.D.

Lemma 8 *Elements of the kernel matrices, $\Sigma_{\boldsymbol{x}}$, are bounded for $\boldsymbol{x}$ in a region I if and only if the eigenvalues of the kernel matrices are bounded in I. Furthermore, if the eigenvalues and the eigenvector elements are sample path differentiable, the matrix elements are sample path differentiable.*

Proof: Consider the spectral decomposition of a kernel matrix $\Sigma_{\boldsymbol{x}} = \Gamma_{\boldsymbol{x}} \Lambda_{\boldsymbol{x}} \Gamma_{\boldsymbol{x}}^T$. Element-wise, and suppressing the dependence on $\boldsymbol{x}$, this gives us

$$\Sigma_{ij} = \sum_k \Gamma_{ik} \Gamma_{jk} \Lambda_{kk}.$$

Since the eigenvectors are normalized vectors, their elements are bounded. Therefore the elements of $\Sigma_{\boldsymbol{x}}$ are bounded if the eigenvalues, Λ_{kk}, are bounded. Equivalently, $\Lambda_{\boldsymbol{x}} = \Gamma_{\boldsymbol{x}}^T \Sigma_{\boldsymbol{x}} \Gamma_{\boldsymbol{x}}$, so the eigenvalues can be expressed as products involving the kernel matrix elements and must be bounded if the matrix elements are bounded. Finally, based on the expansion of $\Sigma_{\boldsymbol{x},ij}$ in terms of sums and products of the eigenvector elements and eigenvalues, it's clear that if the eigenvector elements and eigenvalues are sample path differentiable, the kernel matrix elements will be as well. Q.E.D.

2.5.4 Smoothness properties of isotropic correlation functions

The stationary correlation functions on which I focus in this work are all positive definite on $\Re^p, p = 1, 2, \ldots$, and can therefore be expressed using the Schoenberg (1938) representation (2.6) as a scale mixture of squared exponential correlation functions. In this section I use the representation to show the differentiability properties of the correlation functions and in Section 2.5.5 I do the same for the counterpart nonstationary correlation functions. I show that the smoothness properties of the scale mixture correlation functions are directly related to the existence of moments of the scale parameter.

In the isotropic case, for mean square differentiability, I use the condition on the spectral representation (2.8), to show that if M moments of the scale parameter exist then the Mth-order mean square partial derivatives exist (Section 2.5.4.1). Next, to assess sample path differentiability, I work with the derivatives of the correlation function, as I have not been able to make progress using the spectral representation. Unfortunately, using the derivatives of the correlation function, I am only able to show that the existence of $2M$ moments of the scale parameter is sufficient for Mth-order sample path differentiability (Section 2.5.4.2). In Table 2.1 I give an overview of the smoothness properties of the correlation functions I have been discussing.

2.5.4.1 Mean square differentiability and scale mixtures

Theorem 9 *A stochastic process, $Z(\cdot)$, with isotropic correlation function that can be expressed in the Schoenberg (1938) representation (2.6) has Mth-order mean square partial derivatives if M*

Table 2.1. Smoothness properties of Gaussian processes parameterized by various correlation functions. The asterisk indicates that the sample path part of this statement is a conjecture. In Section 2.5.4.2 I prove only that the Matérn is $\lceil \frac{\nu}{2} - 1 \rceil$ times sample path differentiable

	Smoothness Properties	
	Mean square and sample path	Mean square
Correlation form	differentiability	analyticity
Power exponential, $\nu < 2$	no	no
Matérn	$\lceil \nu - 1 \rceil$ times*	no
Squared exponential	infinitely	yes
Rational quadratic	infinitely	yes

moments of the scale parameter, S, are finite.

Proof: Adler (1981, p. 31) gives the general relationship between the existence of spectral moments and mean square differentiability in the stationary setting. The Mth-order mean square partial derivatives exist if

$$(-1)^M \left. \frac{\partial^{2M} R(\boldsymbol{\tau})}{\partial \tau_{p_1}^2 \cdots \partial \tau_{p_M}^2} \right|_{\boldsymbol{\tau}=\mathbf{0}} = \int_{\Re^P} w_{p_1}^2 \cdots w_{p_M}^2 dH_W(\boldsymbol{w}) < \infty,$$

where $p_m \in \{1, \ldots, P\}, m \in \{1, \ldots, M\}$. So to show the existence of the Mth-order mean square derivative, we need to consider the $2M$th-order moments of the spectral density. Expressed in terms of the Schoenberg (1938) representation (2.6), these are

$$\begin{aligned}
\int_{\Re^P} w_{p_1}^2 \cdots w_{p_M}^2 dH_W(\boldsymbol{w}) &= \int_{\Re^P} w_{p_1}^2 \cdots w_{p_M}^2 h_W(\boldsymbol{w}) d\boldsymbol{w} \\
&= \int_{\Re^P} w_{p_1}^2 \cdots w_{p_M}^2 \int_{\Re^P} \exp(-i\boldsymbol{w}^T \boldsymbol{\tau}) R(\boldsymbol{\tau}) d\boldsymbol{\tau} \\
&\propto \int_{\Re^P} w_{p_1}^2 \cdots w_{p_M}^2 \int_{\Re^P} \int \exp(-i\boldsymbol{w}^T \boldsymbol{\tau}) \exp(-\boldsymbol{\tau}^T \boldsymbol{\tau} s) dH_S(s) d\boldsymbol{\tau} d\boldsymbol{w}.
\end{aligned}$$

First, interchange the order of integration with respect to s and $\boldsymbol{\tau}$ by Fubini's theorem, which is justified because the exponential functions are bounded. Recognize that the integral with respect to $\boldsymbol{\tau}$ is in the form of a normal density with $\boldsymbol{\tau} \sim \mathrm{N}_P\left(-\frac{i\boldsymbol{w}}{2s}, \frac{1}{2s}I\right)$. This gives us

$$\int_{\Re^P} w_{p_1}^2 \cdots w_{p_M}^2 dH_W(\boldsymbol{w}) \propto \int_{\Re^P} \int w_{p_1}^2 \cdots w_{p_M}^2 \left(\frac{1}{2s}\right)^{\frac{P}{2}} \exp\left(-\frac{\boldsymbol{w}^T\boldsymbol{w}}{4s}\right) dH_S(s) d\boldsymbol{w}.$$

Once again interchange the order of integration by Fubini's theorem, which is justified because the integrand is non-negative, and the next steps will show that it is integrable. We see that the integral with respect to $\boldsymbol{w}$ takes the form of a product of moments with respect to $\boldsymbol{w} \sim N_P(0, 2sI)$. A straightforward calculation shows that $E\left(W_{p_1}^2 \cdots W_{p_M}^2\right) \propto S^M$, which gives us

$$\begin{aligned} \int_{\Re^p} w_{p_1}^2 \cdots w_{p_M}^2 dH_W(\boldsymbol{w}) &\propto \int s^M dH_S(s) \\ &= E\left(S^M\right). \end{aligned}$$

So we see that the Mth-order mean square partial derivatives exist if the scale parameter, S, has M moments.

Q.E.D.

2.5.4.2 Sample path differentiability and scale mixtures

I next show that the existence of $2M$ moments of S is sufficient for Mth-order sample path differentiability.

Theorem 10 *A Gaussian process, $Z(\cdot)$, with isotropic correlation function that can be expressed in the Schoenberg (1938) representation (2.6), is Mth-order sample path differentiable if $2M$ moments of the scale parameter, S, are finite.*

Proof: To evaluate the condition (2.10) for $\boldsymbol{x}$ and $\boldsymbol{y}$ in a region I and thereby show continuity of the derivative processes, we need to calculate the covariance functions of the Mth-order derivative processes, which by Adler (1981, p. 27) are of the form

$$(-1)^M \frac{\partial^{2M} R(\boldsymbol{\tau})}{\partial\tau_{p_1}^2 \cdots \partial\tau_{p_M}^2}.$$

If $ES^{2M} < \infty$, we can interchange differentiation and integration when the term inside the integral is of order $S^m, m < 2M$ by lemma 7. The $2M$th partial derivative of a correlation function of the form (2.6) is

$$\begin{aligned}
&c_M \int s^M \exp\left(-\boldsymbol{\tau}^T\boldsymbol{\tau}s\right) dH(s) \\
&+c_{M+1}\left(\sum_{i,j}\tau_{p_i}\tau_{p_j}\right)\int s^{M+1}\exp\left(-\boldsymbol{\tau}^T\boldsymbol{\tau}s\right) dH(s) \\
&+\cdots + c_{2m}\tau_{p_1}^2\cdots\tau_{p_M}^2\int s^{2M}\exp\left(-\boldsymbol{\tau}^T\boldsymbol{\tau}s\right) dH(s) \qquad (2.12)
\end{aligned}$$

for $i, j \in \{1, \ldots, M\}$. To evaluate the condition (2.10) we need to consider the boundedness of

$$(-1)^M \left.\frac{\partial^{2M}R(\boldsymbol{\tau})}{\partial\tau_{p_1}^2\cdots\partial\tau_{p_M}^2}\right|_{\boldsymbol{\tau}=\mathbf{0}} - (-1)^M \left.\frac{\partial^{2M}R(\boldsymbol{\tau})}{\partial\tau_{p_1}^2\cdots\partial\tau_{p_M}^2}\right|_{\boldsymbol{\tau}=\boldsymbol{\tau}_0}. \qquad (2.13)$$

Consider the difference (2.13) of the terms in (2.12). All but one of the differences are bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ in I when multiplied by $|\log\|\boldsymbol{\tau}\||^{1+\epsilon}$ based on the moment condition, boundedness of the exponential function and by lemma 5 since $\tau_{p_m} = O_I(\|\boldsymbol{\tau}\|)$. The exception is the difference of the first term, which is

$$\begin{aligned}
c_M \int s^M \left(1-\exp\left(-\boldsymbol{\tau}^T\boldsymbol{\tau}s\right)\right) dH(s) &= c_M \int s^M s\cdot c_1 \sum_{i,j}\tau_{p_i}\tau_{p_j}\exp(-\boldsymbol{c}^T\boldsymbol{c}s)dH(s) \\
&\propto \sum_{i,j}\tau_{p_i}\tau_{p_j}\int s^{M+1}\exp(-\boldsymbol{c}^T\boldsymbol{c}s)dH(s), \qquad (2.14)
\end{aligned}$$

where the first equation follows by a multivariate Taylor expansion (Schervish 1995, p. 665), and the constant c_1 is a function of $\boldsymbol{c}$, which lies on the line segment joining $\boldsymbol{\tau}$ and $\mathbf{0}$. Again by lemma 5, the moment condition (which is applicable since $M+1 \leq 2M$ for $M \geq 1$) and the boundedness of the exponential function, we have that (2.14), when multiplied by $|\log\|\boldsymbol{\tau}\||^{1+\epsilon}$, is bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ in I. This demonstrates that all of the Mth-order partial derivative processes are sample path continuous based on their covariance function; this gives us sample path differentiability as discussed in Section 2.5.2. Since I was arbitrary, the result holds throughout $\mathcal{X}$.
Q.E.D.

2.5.4.3 Application of results to specific correlation functions

Let's consider how these results on mean square and sample path differentiability apply to the correlation functions considered in this work. The exponential correlation is not mean square differentiable since its scale parameter has an inverse-gamma $\left(\frac{1}{2}, \frac{1}{4}\right)$ distribution (nor is the general power exponential correlation for $\nu < 2$ by seeing that the correlation function is not twice differentiable) and hence is not sample path differentiable by Cambanis (1973, Theorem 6). For the squared exponential and rational quadratic correlation functions, their respective scale parameter distributions are a point mass at 1 and a gamma distribution, both of which have infinitely many moments. Hence Gaussian processes with these correlation functions are infinitely sample path differentiable. The remaining correlation of interest is the Matérn. I have shown that for $\nu > M$, the Matérn gives processes that are at least Mth-order mean square differentiable, while for $\nu > 2M$ the processes are at least Mth-order sample path differentiable.

2.5.4.4 Mean square analyticity

We have seen that both the squared exponential and rational quadratic correlation functions give Gaussian processes with infinitely many mean square and sample path derivatives. I next show that both of these correlation functions are also mean square analytic on $\Re^1$. Loève (1978, p. 137) gives the result that a process is analytic in mean square if and only if the covariance function is analytic at every point (x, x), which for stationary functions simplifies to the point 0. By Churchill (1960, p. 40) a function is analytic at a point if the derivative as a function of a complex argument, $\tau = a + bi$, exists at that point and every point in a neighborhood of the point. Here I show that for processes on $\Re^1$, both the squared exponential and rational quadratic correlation functions are analytic functions at 0. Without loss of generality, let the scale parameter $\kappa = 1$. For the squared exponential correlation function,

$$\frac{\partial}{\partial \tau} \exp(-\tau^2) = -2\tau \exp(-\tau^2)$$

is finite for all complex τ and hence is differentiable in a neighborhood of 0. For the rational quadratic correlation function,

$$\frac{\partial}{\partial \tau} \left(\frac{1}{1+\tau^2} \right)^{\nu} = 2\nu\tau \left(\frac{1}{1+\tau^2} \right)^{\nu-1}$$

is finite for $\tau \neq \pm i$. Hence in a neighborhood of 0, the derivative exists and so the rational quadratic function is analytic at 0. Hence processes with squared exponential or rational quadratic correlation are mean square analytic. While I have not seen proof of the result that processes with these correlation functions are sample path analytic, this may well be the case, which would provide further evidence that they may not be good choices for modelling data (Stein 1999).

2.5.5 Smoothness properties of kernel convolution covariance functions

Here I prove that Gaussian processes with nonstationary correlation of the generalized kernel convolution form,

$$\begin{aligned} R(\boldsymbol{x},\boldsymbol{y}) &= \frac{2^{\frac{p}{2}}\left|\Sigma_{\boldsymbol{x}}\right|^{\frac{1}{4}}\left|\Sigma_{\boldsymbol{y}}\right|^{\frac{1}{4}}}{\left|\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{y}}\right|^{\frac{1}{2}}} R\left(\sqrt{Q_{\boldsymbol{xy}}}\right) \\ &= \frac{2^{\frac{p}{2}}\left|\Sigma_{\boldsymbol{x}}\right|^{\frac{1}{4}}\left|\Sigma_{\boldsymbol{y}}\right|^{\frac{1}{4}}}{\left|\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{y}}\right|^{\frac{1}{2}}} \int_0^\infty \exp\left(-Q_{\boldsymbol{xy}}s\right) dH(s), \end{aligned} \tag{2.15}$$

where

$$Q_{\boldsymbol{xy}} = (\boldsymbol{x}-\boldsymbol{y})^T \left(\frac{\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{y}}}{2}\right)^{-1} (\boldsymbol{x}-\boldsymbol{y}), \tag{2.16}$$

have smoothness properties similar to those of the isotropic correlation function, $R(\tau)$, on which they are based, provided the underlying kernel matrices vary smoothly. This result makes sense intuitively because if the kernels are smooth, then in a small neighborhood, the covariance is nearly stationary, so the smoothness properties should depend on the properties of the underlying isotropic correlation. The anisotropy in the correlation due to $\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{y}}$ will not play a role in smoothness properties, because this merely rotates and scales the space. This does not affect smoothness properties, which relate to the behavior of a stationary covariance function near the origin.

The nonstationary correlation functions (2.15) are scale mixtures of the HSK nonstationary correlation (2.3). For these nonstationary scale mixtures, just as I demonstrated for the stationary scale mixtures, differentiability properties are directly related to existence of moments of the scale parameter. Here I prove that if $ES^{2M} < \infty$, then stochastic processes with the generalized kernel convolution covariance are Mth-order mean square differentiable (Section 2.5.5.1), and Gaussian processes are Mth-order sample path differentiable (Section 2.5.5.2).

2.5.5.1 Mean square differentiability

Theorem 11 *A nonstationary stochastic process, $Z(\cdot)$, has Mth-order mean square derivatives if it has a correlation function that can be expressed in the form (2.15), the scale parameter, S, has*

$2M$ moments, the elements of the kernel matrices, Σ_x, are M times sample path differentiable, and the kernel matrices are not singular.

Proof:

By Adler (1981, p. 27), the finiteness of

$$\left.\frac{\partial^{2M} R(\boldsymbol{x},\boldsymbol{y})}{\partial x_{p_1}\cdots\partial x_{p_M}\partial y_{p_1}\cdots\partial y_{p_M}}\right|_{\boldsymbol{x}=\boldsymbol{y}=\boldsymbol{u}}$$

for $p_m \in \{1,\ldots,P\}, m \in \{1,\ldots,M\}$ is sufficient for the existence of Mth-order mean square partial derivative processes. The $2M$th partial derivatives of (2.15) take the form

$$\sum_{m_1=0}^{M}\sum_{m_2=0}^{M}\sum_{m_3=0}^{2M}\sum_{m_4=0}^{2M} D_{\boldsymbol{x}}^{(m_1)}\left(|\Sigma_x|^{\frac{1}{4}}\right) D_{\boldsymbol{y}}^{(m_2)}\left(|\Sigma_y|^{\frac{1}{4}}\right)$$
$$\times D^{(m_3)}\left(|\Sigma_x+\Sigma_y|^{-\frac{1}{2}}\right) D^{(m_4)}\left(\int \exp(-Q_{xy}s)dH(s)\right). \tag{2.17}$$

First consider $D_{\boldsymbol{x}}^{(m_1)}\left(|\Sigma_{\boldsymbol{x}}|^{\frac{1}{4}}\right)$. The highest order derivative involving only terms in $\boldsymbol{x}$ is of the form $D_{\boldsymbol{x}}^{(M)}\left(|\Sigma_{\boldsymbol{x}}|^{\frac{1}{4}}\right)$. By assumption, the kernel matrices are not singular, so negative powers after differentiation do not cause the expression to be infinite. A determinant can be expressed as a product of the elements in the matrix. By assumption, the elements of $\Sigma_{\boldsymbol{x}}$ are M times sample path differentiable, so the Mth-order derivative of the determinant is finite. The same argument holds by symmetry for $D_{\boldsymbol{y}}^{(m_2)}\left(|\Sigma_{\boldsymbol{y}}|^{\frac{1}{4}}\right)$. Next consider $D^{(m_3)}\left(|\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{y}}|^{\frac{1}{4}}\right)$. Once again, since the matrices are assumed to not be singular, the power does not pose a problem. The determinant can be considered as the product of the sum of elements of $\Sigma_{\boldsymbol{x}}$ and $\Sigma_{\boldsymbol{y}}$. By assumption the Mth-order derivatives with respect to $\boldsymbol{x}$ and $\boldsymbol{y}$ exist, so once again, the derivative is finite.

The final term is $D^{(m_4)}\left(\int \exp(-Q_{xy}s)dH(s)\right)$. By lemma 7, we can interchange differentiation and integration. Since finiteness of the $2M$th order derivatives implies finiteness of lower order derivatives, we need only assess

$$\begin{aligned} D^{(2M)}\left(\int \exp(-Q_{xy}s)dH(s)\right) &= c_1 D_{x_{p_1}}^{(1)}(Q_{\boldsymbol{xy}})\cdots D_{y_{p_M}}^{(1)}(Q_{\boldsymbol{xy}})\int s^{2M}\exp(-Q_{xy}s)dH(s) \\ &\quad + \ldots + c_{2M} D^{2M} Q_{xy}\int s\exp(-Q_{xy}s)dH(s). \end{aligned}$$

First consider the integrals. Using the assumption on the moments of S and the boundedness of the exponential function, the integrals are bounded. By lemma 6 the derivatives of Q_{xy} are $O_I(1)$, so they are bounded.

I have shown that all the terms in the $2M$th-order partial derivative (2.17) of the correlation function are finite, and therefore that the $2M$th-order mean square derivative exists. Since the argument applies to arbitrary $2M$th-order partial derivatives of the correlation and holds for arbitrary $\boldsymbol{u}$, all $2M$-th order mean square derivatives exist on $\mathcal{X}$.
Q.E.D.

2.5.5.2 Sample path differentiability

Theorem 12 *A nonstationary Gaussian process, $Z(\cdot)$, is Mth-order sample path differentiable if its correlation function can be expressed in the form (2.15), the scale parameter, S, has $2M$ moments, the elements of the kernel matrices, Σ_x, are $M+1$ times sample path differentiable, and the kernel matrices are not singular.*

Proof: To assess the condition (2.9) for $\boldsymbol{x}$ and $\boldsymbol{y}$ in a region I, we need to assess $E|Z(\boldsymbol{x})-Z(\boldsymbol{y})|^2 = (E(Z(\boldsymbol{x})^2) - E(Z(\boldsymbol{x})Z(\boldsymbol{y}))) + (E(Z(\boldsymbol{y})^2 - E(Z(\boldsymbol{x})Z(\boldsymbol{y})))$. By symmetry, we need to consider only one of the two terms:

$$\begin{aligned} E(Z(\boldsymbol{x})^2) - E(Z(\boldsymbol{x})Z(\boldsymbol{y})) &= C(\boldsymbol{x},\boldsymbol{x}) - C(\boldsymbol{x},\boldsymbol{y}) \\ &= D^{(2M)}(R(\boldsymbol{x},\boldsymbol{y}))|_{\boldsymbol{x},\boldsymbol{x}} - D^{(2M)}(R(\boldsymbol{x},\boldsymbol{y}))|_{\boldsymbol{x},\boldsymbol{y}}. \end{aligned} \quad (2.18)$$

The $2M$th-order partial derivatives of the nonstationary correlation (2.15) are of the form

$$\begin{aligned} D^{(2M)}R(x,y) &= \sum_{m_1=0}^{M}\sum_{m_2=0}^{M}\sum_{m_3=0}^{2M}\sum_{m_4=0}^{2M} D_{\boldsymbol{x}}^{(m_1)}\left(|\Sigma_{\boldsymbol{x}}|^{\frac{1}{4}}\right) D_{\boldsymbol{y}}^{(m_2)}\left(|\Sigma_{\boldsymbol{y}}|^{\frac{1}{4}}\right) \\ &\quad \times D^{(m_3)}\left(|\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{y}}|^{-\frac{1}{2}}\right) D^{(m_4)}\left(\int \exp(-Q_{\boldsymbol{xy}}s)dH(s)\right), \end{aligned} \quad (2.19)$$

with the constraint that $m_1+m_2+m_3+m_4 = 2M$. Next, let

$$\begin{aligned} g_1(\boldsymbol{x},\boldsymbol{y}) &= D_{\boldsymbol{x}}^{(m_1)}\left(|\Sigma_{\boldsymbol{x}}|^{\frac{1}{4}}\right) \\ g_2(\boldsymbol{x},\boldsymbol{y}) &= D_{\boldsymbol{y}}^{(m_2)}\left(|\Sigma_{\boldsymbol{y}}|^{\frac{1}{4}}\right) \\ g_3(\boldsymbol{x},\boldsymbol{y}) &= D^{(m_3)}\left(|\Sigma_{\boldsymbol{x}}+\Sigma_{\boldsymbol{y}}|^{-\frac{1}{2}}\right) \\ g_4(\boldsymbol{x},\boldsymbol{y}) &= D^{(m_4)}\left(\int \exp(-Q_{\boldsymbol{xy}}s)dH(s)\right), \end{aligned}$$

and successively apply the identity,

$$f(\boldsymbol{x},\boldsymbol{x})g(\boldsymbol{x},\boldsymbol{x}) - f(\boldsymbol{x},\boldsymbol{y})g(\boldsymbol{x},\boldsymbol{y}) = (f(\boldsymbol{x},\boldsymbol{x}) - f(\boldsymbol{x},\boldsymbol{y}))g(\boldsymbol{x},\boldsymbol{x}) - (g(\boldsymbol{x},\boldsymbol{x}) - g(\boldsymbol{x},\boldsymbol{y}))g(\boldsymbol{x},\boldsymbol{y}), \tag{2.20}$$

to

$$g_1(\boldsymbol{x},\boldsymbol{x})g_2(\boldsymbol{x},\boldsymbol{x})g_3(\boldsymbol{x},\boldsymbol{x})g_4(\boldsymbol{x},\boldsymbol{x}) - g_1(\boldsymbol{x},\boldsymbol{y})g_2(\boldsymbol{x},\boldsymbol{y})g_3(\boldsymbol{x},\boldsymbol{y})g_4(\boldsymbol{x},\boldsymbol{y}),$$

i.e., one of the terms in the sum (2.19). This gives us that (2.18) can be expressed as

$$\begin{aligned}\sum_{m_1=0}^{M}\sum_{m_2=0}^{M}\sum_{m_3=0}^{2M}\sum_{m_4=0}^{2M} \quad & \left(g_1(\boldsymbol{x},\boldsymbol{x}) - g_1(\boldsymbol{x},\boldsymbol{y})\right) g_2(\boldsymbol{x},\boldsymbol{x})g_3(\boldsymbol{x},\boldsymbol{x})g_4(\boldsymbol{x},\boldsymbol{x}) \\ & + \left(g_2(\boldsymbol{x},\boldsymbol{x}) - g_2(\boldsymbol{x},\boldsymbol{y})\right) g_1(\boldsymbol{x},\boldsymbol{y})g_3(\boldsymbol{x},\boldsymbol{x})g_4(\boldsymbol{x},\boldsymbol{x}) \\ & + \left(g_3(\boldsymbol{x},\boldsymbol{x}) - g_3(\boldsymbol{x},\boldsymbol{y})\right) g_1(\boldsymbol{x},\boldsymbol{y})g_2(\boldsymbol{x},\boldsymbol{y})g_4(\boldsymbol{x},\boldsymbol{x}) \\ & + \left(g_4(\boldsymbol{x},\boldsymbol{x}) - g_4(\boldsymbol{x},\boldsymbol{y})\right) g_1(\boldsymbol{x},\boldsymbol{y})g_2(\boldsymbol{x},\boldsymbol{y})g_3(\boldsymbol{x},\boldsymbol{y}).\end{aligned} \tag{2.21}$$

To satisfy the condition (2.9) I need only show that for $i \in \{1,2,3,4\}$ $(g_i(\boldsymbol{x},\boldsymbol{x}) - g_i(\boldsymbol{x},\boldsymbol{y}))|\log\|\boldsymbol{x} - \boldsymbol{y}\||^{1+\epsilon}$ is bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ in I. By lemma 5 it is sufficient that $(g_i(\boldsymbol{x},\boldsymbol{x}) - g_i(\boldsymbol{x},\boldsymbol{y})) = O_I(\|\boldsymbol{x} - \boldsymbol{y}\|)$, which is satisfied by lemma 4 if $g_i(\boldsymbol{x},\boldsymbol{y})$ is once differentiable. $g_i(\boldsymbol{x},\boldsymbol{y})$ itself will involve at most M derivatives with respect to each of $\boldsymbol{x}$ and $\boldsymbol{y}$, so satisfying lemma 4 will involve at most the existence of $M+1$ derivatives; I focus on the highest order derivatives, as the other order derivatives are differentiable if the highest order derivatives are differentiable.

For $g_1(\boldsymbol{x},\boldsymbol{y})$ (equivalently for $g_2(\boldsymbol{x},\boldsymbol{y})$) the determinant can be expressed as a product of the elements of $\Sigma_{\boldsymbol{x}}$, so $g_1(\boldsymbol{x},\boldsymbol{y})$ is differentiable by the assumption that the elements of the kernel matrices are $M+1$ times differentiable. Note that raising the determinant to a power has no effect because the kernel matrices are assumed to not be singular. Similarly, for $g_3(\boldsymbol{x},\boldsymbol{y})$, the determinant can be considered as a product of terms of $\Sigma_{\boldsymbol{x}} + \Sigma_{\boldsymbol{y}}$, so $g_3(\boldsymbol{x},\boldsymbol{y})$ is differentiable by the assumption that the elements of the kernel matrices are $M+1$ times differentiable (since the $2M$th-order partial derivative involves at most M partial derivatives with respect to each of $\boldsymbol{x}$ and $\boldsymbol{y}$). Next consider the $2M$th-order partial derivative of $g_4(\boldsymbol{x},\boldsymbol{y})$,

$$\begin{aligned}D^{(2M)}\left(\int \exp(-Q_{\boldsymbol{xy}}s)dH(s)\right) \quad = \quad & D^{(2M)}(Q_{\boldsymbol{xy}})\int s\exp(-Q_{\boldsymbol{xy}}s)dH(s) \\ & + \cdots + D^{(1)}_{x_1}(Q_{\boldsymbol{xy}})\cdots D^{(1)}_{y_M}(Q_{\boldsymbol{xy}}) \\ & \times \int s^{2M}\exp(-Q_{\boldsymbol{xy}}s)dH(s),\end{aligned} \tag{2.22}$$

where I interchange differentiation and integration based on lemma 7. First consider the difference

$$D^{(1)}_{x_1}(Q_{\boldsymbol{x}\boldsymbol{y}})\cdots D^{(1)}_{y_M}(Q_{\boldsymbol{x}\boldsymbol{y}})\int s^{2M}\exp(-Q_{\boldsymbol{x}\boldsymbol{y}}s)dH(s)\Big|_{\boldsymbol{x},\boldsymbol{x}}$$
$$-D^{(1)}_{x_1}(Q_{\boldsymbol{x}\boldsymbol{y}})\cdots D^{(1)}_{y_M}(Q_{\boldsymbol{x}\boldsymbol{y}})\int s^{2M}\exp(-Q_{\boldsymbol{x}\boldsymbol{y}}s)dH(s)\Big|_{\boldsymbol{x},\boldsymbol{y}}.$$

$D^{(1)}(Q_{\boldsymbol{x}\boldsymbol{y}})\Big|_{\boldsymbol{x},\boldsymbol{x}} = 0$ and in the second term $D^{(1)}(Q_{\boldsymbol{x}\boldsymbol{y}})\Big|_{\boldsymbol{x},\boldsymbol{y}} = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$ by lemma 6, so the whole expression is $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$. Note that I make use of the existence of the $2M$th moment of S and the boundedness of the exponential term. Hence this difference satisfies (2.9). Next consider the remaining terms in (2.22). Let

$$\begin{aligned} g_5(\boldsymbol{x},\boldsymbol{y}) &= D^{(m_5)}(Q_{\boldsymbol{x}\boldsymbol{y}}) \\ g_6(\boldsymbol{x},\boldsymbol{y}) &= \int s^{m_6}\exp(-Q_{\boldsymbol{x}\boldsymbol{y}}s)dH(s), \end{aligned}$$

where $m_6 \le 2M-1$ and $D^{(m_5)}(Q_{\boldsymbol{x}\boldsymbol{y}})$ is a product of terms involving derivatives of various orders of $Q_{\boldsymbol{x}\boldsymbol{y}}$ with $m_5 \in \{1,\ldots,2M\}$. Applying the identity (2.20) to $g_5(\boldsymbol{x},\boldsymbol{x})g_6(\boldsymbol{x},\boldsymbol{x}) - g_5(\boldsymbol{x},\boldsymbol{y})g_6(\boldsymbol{x},\boldsymbol{y})$ I need only show that $g_5(\boldsymbol{x},\boldsymbol{y})$ and $g_6(\boldsymbol{x},\boldsymbol{y})$ are once differentiable to satisfy (2.9). First consider differentiating $g_5(\boldsymbol{x},\boldsymbol{y})$. The derivative $D^{(m_1)}(Q_{\boldsymbol{x}\boldsymbol{y}})$ is at most order $2M$ and therefore order M in either $\boldsymbol{x}$ or $\boldsymbol{y}$. Since I have assumed $M+1$ derivatives of the kernel matrix elements, by lemma 4, $g_5(\boldsymbol{x},\boldsymbol{x})-g_5(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$. Next I show that $g_6(\boldsymbol{x},\boldsymbol{x})-g_6(\boldsymbol{x},\boldsymbol{y})$ is at least $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$ using a multivariate Taylor expansion (Schervish 1995, p. 665) of $\exp(-Q_{\boldsymbol{x}\boldsymbol{y}}s)$ at $\boldsymbol{y}=\boldsymbol{x}$ with first-order remainder:

$$1-\exp(-Q_{\boldsymbol{x}\boldsymbol{y}}s) \propto \sum_p sD^{(1)}_{y_p}(Q_{\boldsymbol{x}\boldsymbol{y}})|_{\boldsymbol{x},\boldsymbol{c}}\exp(-Q_{\boldsymbol{x}\boldsymbol{c}}s)(x_p-y_p) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)\cdot s,$$

where $\boldsymbol{c}$ lies on the line segment joining $\boldsymbol{x}$ and $\boldsymbol{y}$. So we have

$$g_6(\boldsymbol{x},\boldsymbol{x})-g_6(\boldsymbol{x},\boldsymbol{y}) = O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)c_1\int s^{m_2+1}\exp(-Q_{\boldsymbol{x}\boldsymbol{c}}s)dH(s),$$

and since $m_2+1 \le 2M$, the integral is bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ on I, and hence the whole expression is $O_I(\|\boldsymbol{x}-\boldsymbol{y}\|)$.

Therefore, all terms in (2.21) are bounded for $\boldsymbol{x}$ and $\boldsymbol{y}$ on I when divided by $\|\boldsymbol{x}-\boldsymbol{y}\|$ and the condition (2.9) is satisfied by lemma 5. Since I was arbitrary, the result holds throughout $\mathcal{X}$.

Q.E.D.

2.5.5.3 Implications for nonstationary modelling

For the squared exponential and rational quadratic nonstationary correlation functions, which have infinitely many moments of the scale parameter, Gaussian process sample paths are infinitely mean square and sample path differentiable given sufficient smoothness in the kernel matrices. For the Matérn nonstationary correlation function with $\nu > 2M$, M mean square and sample path derivatives are guaranteed.

In constructing models, the elements of the kernels, $\Sigma_{\boldsymbol{x}}$, will themselves be random fields. It is sufficient that these elements (or the eigenvalues and eigenvector elements by lemma 8) be sample path differentiable to the Mth or ($M+1$)th order to have Mth-order mean square or sample path differentiability, respectively. This suggests that at the highest level in the model hierarchy, one will need a stationary covariance structure to easily guarantee the desired sample path differentiability. In the regression modelling of Chapter 4, this stationarity is imposed on the kernel eigenstructure.

Also note that to use a nonstationary covariance model, as opposed to a correlation model, one introduces a variance function $\sigma^2(\boldsymbol{x})$. In my applications, I take this to be constant and therefore it has no effect on differentiability of the resulting processes. However, one could easily take this to be a random field and if it has $M+1$ sample path derivatives, it is easy to show that theorems 11 and 12 continue to hold using arguments analogous to those regarding differentiation of the determinants of the kernels.

2.6 Discussion

This chapter introduces a class of nonstationary correlation functions based on familiar stationary correlation functions. By extending the original Higdon et al. (1999) kernel convolution method, the class provides a much broader set of nonstationary correlation functions than previously available. Some of these correlation functions may be better able to model particular datasets than the nonstationary correlation based on the squared exponential form, because they are more flexible in various respects.

In particular, I have provided nonstationary correlation functions with a range of smoothness properties, in contrast to the original HSK nonstationary covariance function (Fuentes and Smith

2001). The HSK approach gives infinitely differentiable sample paths unless a lack of smoothness is enforced through the kernel structure, which would be a rather ad hoc way to reduce smoothness. One of the new functions is a nonstationary version of the attractive Matérn correlation function, which has a parameter that indexes the mean square and sample path differentiability of Gaussian processes with this correlation function. With the new nonstationary correlation functions, one can create a smooth underlying kernel structure and yet retain control over sample path smoothness. In addition, one can create one's own nonstationary correlation function using any distribution for the scale parameter, S, involved in the generalization of HSK.

A related advantage of the generalization involves asymptotic behavior. Stein (1999) has shown that the convergence of kriging methods relies on the behavior of the correlation function near the origin and the compatibility, in a certain technical sense, of the modelled correlation function with the true correlation. He recommends the Matérn correlation because of its ability to adapt to different behavior near the origin, while the power exponential family is restricted to the extremes of non-differentiability and analytic behavior. Hence the nonstationary Matérn correlation function introduced in this chapter may be of particular interest.

The results in this chapter give sufficient, but seemingly not necessary conditions for smoothness properties of stochastic processes. In the stationary case, I have proven that the existence of $2M$ moments of the scale parameter is sufficient for Mth-order sample path differentiability. I suspect the condition can be weakened so that the existence of M moments is equivalent to Mth-order sample path differentiability. In the case of the Matérn, which is a mixture of the squared exponential correlation with a scale parameter distributed as inverse-gamma $\left(\nu, \frac{1}{4}\right)$ ($\lceil \nu - 1 \rceil$ moments), this would give exactly Mth-order sample path differentiability when $M < \nu \leq M + 1$, which is the same condition as for mean square differentiability (Stein 1999, p. 32). Furthermore, based on Cambanis (1973, Theorem 6), we know that the Matérn has no more than M sample path derivatives when $M < \nu \leq M + 1$. For the special case of the Matérn , one may be able to prove the sample path differentiability result by representing the correlation function as an infinite sum based on the properties of Bessel functions, for which Stein (1999, p. 32) and references therein would be a starting point. I do not have any suggestions for the general scale mixture case. For mean square and sample path differentiability of the nonstationary kernel convolution correlation functions, it may be possible to weaken the $2M$ moment condition to the existence of M moments of the scale parameter, although I do not have a suggestion for how to proceed. Also, in the case of sample path differentiability, I have required $M + 1$ sample path derivatives of the kernel matrix elements for ease of argument, but more subtle reasoning may allow this condition to be weakened to M derivatives.

The exact smoothness conditions are an issue only for the results of this chapter with respect to the Matérn correlation function, since the exponential, rational quadratic, and squared exponential lie in the extremes of differentiability. However, even for the Matérn, the key fact is that the differentiability varies with ν, not the exact number of derivatives as a function of ν, so sharpening the results in this chapter is of limited practical import.