CBMS Lecture 2

Alan E. Gelfand
Duke University

What we will do here

- A brief review of the basic theory of stochastic processes needed for the development of point-referenced spatial data or geo-statistical models.
- Spatial stochastic processes built from independent increment processes, with particular interest in stationary spatial processes.
- The connection between covariance functions and spectral measures.
- The validity of covariance functions.
- Smoothness of process realizations as driven by stationary covariance functions and directional derivative processes
- Nonstationary covariance specifications.

Formal modeling theory for spatial processes

- The collection of random variables $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ or more generally $\{Y(\mathbf{s}) : \mathbf{s} \in \Re^r\}$, envisions a stochastic process indexed by $\mathbf{s}$.
- To capture spatial association, these variables will be pairwise dependent with strength of dependence that is specified by their locations.
- We have to "determine" the joint distribution for an uncountable number of random variables.
- We do this through specification of arbitrary finite dimensional distributions, i.e., for an arbitrary number of and choice of locations.

cont

- This characterizes the stochastic process. More precisely, for the set of locations, $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$, let the finite dimensional distribution be $P(Y(\mathbf{s}_1) \in A_1, Y(\mathbf{s}_2) \in A_2, ..., Y(\mathbf{s}_n) \in A_n)$.
- Two "consistency" conditions (Kolmogorov - iff):
  (i) Under any permutation α of the indices $1, 2, ..., n$, say $\alpha_1, \alpha_2, ..., \alpha_n$, $P(Y(\mathbf{s}_{\alpha_1}) \in A_{\alpha_1}, Y(\mathbf{s}_{\alpha_2}) \in A_{\alpha_2}, ..., Y(\mathbf{s}_{\alpha_n}) \in A_n) = P(Y(\mathbf{s}_1) \in A_1, Y(\mathbf{s}_2) \in A_2, ..., Y(\mathbf{s}_n) \in A_n)$. That is, permutation of the indices does not change the probability of events.
  (ii) For any set of locations, $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$ consider an additional arbitrary location s_{n+1}. If we marginalize the $n+1$ dimensional joint distribution specified for $Y(\mathbf{s}_1), Y(\mathbf{s}_2), ..., Y(\mathbf{s}_n), Y(\mathbf{s}_{n+1})$ over $Y(\mathbf{s}_{n+1})$, we obtain the n dimensional joint distribution specified for $Y(\mathbf{s}_1), Y(\mathbf{s}_2), ..., Y(\mathbf{s}_n)$.

cont.

- Characterizing the entire collection of finite dimensional distributions can be challenging.
- So, Gaussian processes (possibly transformations of) or to mixtures of such processes (a very rich class).
- We can work with multivariate normal distributions
- Only require a mean surface, $\mu(\mathbf{s})$ and a valid correlation function which provides the covariance matrix

cont.

- **INFERENCE:** We only observe $Y(\mathbf{s})$ at a finite set of locations, $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$. Based upon $\{Y(\mathbf{s}_i), i = 1, \ldots, n\}$, we seek to infer about the mean, variability, and association structure of the process. We also seek to predict $Y(\mathbf{s})$ at arbitrary unobserved locations.
- **COMMENT:** Our focus is on hierarchical modeling where the spatial process is introduced through random effects at the second stage of the modeling. Now, the process is latent and the data, modeled at the first stage, helps us to learn about it
- Connections to hierarchical modeling in general and latent variables, e.g., dynamic models

Asymptotics

- A critical difference between the one-dimensional time domain and the two-dimensional spatial domain: we have full order in the former, but only partial order in two or more dimensions.
- **IMPLICATIONS** Asymptotics for time series lets time go to ∞; envisions an increasing time domain.
- Asymptotics for spatial process data envisions a fixed region with more and more points filling in this domain (infilling)
- Increasing domain asymptotics assumes that, as we collect more and more data, we can learn about temporal association at increasing distance in time.
- With infill asymptotics, for a fixed domain, can learn more and more about association as distance between points tends to 0. However, with a maximum distance fixed by the domain cannot learn about association (consistent inference) at increasing distance.
- So, an increasingly better job with regard to spatial prediction at a given location.

Asymptotics

- However, not in terms of inferring about other features of the process.
- Learning about process parameters will be bounded; Fisher information does not go to ∞, Cramèr-Rao lower bounds and asymptotic variances do not go to 0.
- Encourages using a Bayesian framework for inference to avoid asymptotic theory
- However, implies that the data never overwhelms the prior; there is no free lunch

Some basic stochastic process theory for spatial processes

- **REMARK:** When we develop a spatial stochastic process model, we can proceed along two paths.
- We can specify the process stochastically and obtain its induced covariance function, i.e., its induced dependence structure.
- We can start with say, a Gaussian process and then specify a valid covariance function to overlay on the Gaussian process.
- Modeling customarily proceeds from the latter. Here, to develop some theory, we follow the former

Some basic stochastic process theory for spatial processes

- We start with independent increment processes.
- A real-valued independent increment process, Z over say R^2 is such that, for disjoint sets A and B, $Z(A)$ and $Z(B)$ are independent.
- Let $Z(d\mathbf{w})$ be the generator for these random variables in the sense that $Z(A) = \int_A Z(d\mathbf{w})$.
- So, if A and B are disjoint, then $Z(A \bigcup B) = Z(A) + Z(B)$.
- Assume that the Z process has first and second moments and, for convenience that the first moment is 0.
- Set $E(Z^2(A)) = \text{var}(A) \equiv G(A)$, i.e., $G(A) = \int_A G(dw)$. Also, $E(Z(A)Z(B)) = E(Z^2(A \bigcap B)) = G(A \bigcap B)$ due to independence of increments.

Some basic stochastic process theory for spatial processes

- Let $Z_f = \int f(\mathbf{w}) Z(d\mathbf{w})$ for a measurable function f. $Z(A) = \int 1(\mathbf{w} \in A) Z(d\mathbf{w})$ is a special case.
- As usual, use step functions to study the behavior of Z_f for measurable f and the dependence between say Z_{f1} and Z_{f2}.
- If $f(\mathbf{w}) = \sum_l a_l 1(\mathbf{w} \in A_l)$, then $Z_f = \sum_l a_l Z(A_l)$. Because of the independent increments, $\mathrm{var} Z_f = \sum a_l^2 G(A_l)$.
- Consider $f_1(\mathbf{w}) = \sum_l a_l 1(\mathbf{w} \in A_l)$ and $f_2(\mathbf{w}) = \sum_k b_k 1(\mathbf{w} \in B_k)$. Then, $\mathrm{cov}(Z_{f1}, Z_{f2}) =$

$$E(Z_{f1} Z_{f2}) = \sum_l \sum_k a_l b_k E(Z(A_l) Z(B_k)$$

$$= \sum_l \sum_k a_l b_k E(Z^2(A_l \bigcap B_k)) = \sum_l \sum_k a_l b_k G(A_l \bigcap B_k) \, .$$

- But, also, $\int f_1(\mathbf{w}) f_2(\mathbf{w}) G(d\mathbf{w}) = \sum_l \sum_k a_l b_k G(A_l \bigcap B_k)$.
- So, we have that $E(Z_{f1} Z_{f2}) = \int f_1(\mathbf{w}) f_2(\mathbf{w}) G(d\mathbf{w})$.

Some basic stochastic process theory for spatial processes

- Now, bring in the spatial process, defining

$$Y(\mathbf{s}) = \int \psi(\mathbf{s}, \mathbf{w}) Z(d\mathbf{w}) \ ,$$

  i.e., $Y(\mathbf{s})$ is $Z_{\psi_\mathbf{s}}$ in our notation above.
- We allow ψ to be a complex valued function in order to employ the form $\psi(\mathbf{s}, \mathbf{w}) = e^{i\mathbf{s}^T\mathbf{w}}$.
- We also allow Z to be complex valued, introducing the complex conjugate (using an overline).
- So, $G(A) = E(Z(A)\overline{Z}(A)) = E|Z(A)|^2$.
- We have defined a stochastic process and to calculate $\mathrm{cov} Y(\mathbf{s}), Y(\mathbf{s}')$, we compute $E(Y(\mathbf{s})\overline{Y}(\mathbf{s}')) = \int \psi(\mathbf{s}, \mathbf{w})\overline{\psi}(\mathbf{s}', \mathbf{w}) G(d\mathbf{w})$

Some basic stochastic process theory for spatial processes

- With $\psi(\mathbf{s}, \mathbf{w}) = e^{i\mathbf{s}^T\mathbf{w}}$, we obtain $\mathrm{cov}\, Y(\mathbf{s}), Y(\mathbf{s}') = \int e^{i\mathbf{s}^T\mathbf{w}} e^{-i\mathbf{s}'^T\mathbf{w}} G(d\mathbf{w}) = \int e^{i(\mathbf{s}-\mathbf{s}')^T\mathbf{w}} G(d\mathbf{w})$
- The association between $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ depends only upon the separation vector $\mathbf{h} = \mathbf{s} - \mathbf{s}'$, $\mathrm{cov}(Y(\mathbf{s}), Y(\mathbf{s}')) = C(\mathbf{s} - \mathbf{s}')$, STATIONARITY.
- An elegant result (Yaglom) says that $Y(\mathbf{s})$ is a stationary stochastic process if and only if it can be represented in the form $Y(\mathbf{s}) = \int e^{i\mathbf{s}^T\mathbf{w}} Z(d\mathbf{w})$ where Z is a possibly complex-valued, mean 0, independent increments process
- We note the parallel structure, $Y(\mathbf{s}) = \int e^{i\mathbf{s}^T\mathbf{w}} Z(d\mathbf{w})$ and $C(\mathbf{s}) = \int e^{i\mathbf{s}^T\mathbf{w}} G(d\mathbf{w})$.

Some basic stochastic process theory for spatial processes

- Other choices for ψ appear in the literature.
- For instance, let ψ be a kernel function $K(\mathbf{s}-\mathbf{w})$ which is integrable over $\Re^2$.
- Then, $Y(\mathbf{s}) = \int K(\mathbf{s}-\mathbf{w})Z(d\mathbf{w})$ and $\text{cov}(Y(\mathbf{s})Y(\mathbf{s}') = \int K(\mathbf{s}-\mathbf{w})K(\mathbf{s}'-\mathbf{w})G(d\mathbf{w})$.
- If $G(d\mathbf{w}) = \sigma^2 d\mathbf{w}$, after a change of variable, $\text{cov}(Y(\mathbf{s})Y(\mathbf{s}') = \sigma^2 \int K(\mathbf{s}-\mathbf{s}'+\mathbf{u})K(\mathbf{u})d\mathbf{u}$. Again, stationarity
- Such a process construction is called *kernel convolution*
- How rich is the class of stationary process obtainable under kernel convolution?
- When does the foregoing provide a Gaussian process?
- If $Z(\mathbf{s})$ is Brownian motion then $Y(\mathbf{s})$ is a Gaussian process.
- Recall, Brownian motion is an independent increments process providing jointly normally distributed random variables.

Covariance functions and spectra

- To specify a stationary process we must provide a valid covariance function. We need $C(\mathbf{h}) \equiv cov(Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h}))$ to be a positive definite function.
- Verifying the positive definiteness condition is not routine
- Again, *Bochner's Theorem* which provides a necessary and sufficient condition for $C(\mathbf{h})$ to be positive definite.
  $C(\mathbf{h})$ is positive definite if and only if

$$C(\mathbf{h}) = \int \cos(\mathbf{w}^T\mathbf{h})\, G(d\mathbf{w}) \,,$$

  G is a bounded, positive, symmetric about 0 measure in $\Re^2$.
- Then $C(\mathbf{0}) = \int G d(\mathbf{w})$ becomes a normalizing constant, and $G(d\mathbf{w})/C(\mathbf{0})$ is referred to as the *spectral distribution* which induces $C(\mathbf{h})$.
- If $G(d\mathbf{w})$ has a density with respect to Lebesgue measure, i.e., $G(d\mathbf{w}) = g(\mathbf{w})d\mathbf{w}$, then $g(\mathbf{w})/C(\mathbf{0})$ is referred to as the *spectral density*

Covariance functions and spectra

- In theory, Bochner's Theorem can be used to generate valid covariance functions; however integral in closed form only in very special cases
- Since $e^{i\mathbf{w}^T\mathbf{h}} = \cos(\mathbf{w}^T\mathbf{h}) + i\sin(\mathbf{w}^T\mathbf{h})$, we have $C(\mathbf{h}) = \int e^{i\mathbf{w}^T\mathbf{h}} G(d\mathbf{w})$.
- Imaginary term disappears due to the symmetry of G around 0.
- So, $C(\mathbf{h})$ is a valid covariance function iff if it is the characteristic function of a random variable with a symmetric distribution about 0.

Covariance functions and spectra

- The Fourier transform of $C(\mathbf{h})$ is

$$\widehat{c}(\mathbf{w}) = \int e^{-i\mathbf{w}^T\mathbf{h}}\, C(\mathbf{h}) d\mathbf{h}\,.$$

- Applying the inversion formula, $C(\mathbf{h}) = (2\pi)^{-r} \int e^{i\mathbf{w}^T\mathbf{h}} \widehat{c}(\mathbf{w}) d\mathbf{w}$, we see that $(2\pi)^{-r}\widehat{c}(\mathbf{w})/C(0) = g(\mathbf{w})$, the spectral density.
- Explicit computation is usually not possible except in special cases.
- In theory, this can be used to check whether a given $C(\mathbf{h})$ is valid: we simply compute $\widehat{c}(\mathbf{w})$ and check whether it is positive and integrable (so it is a density up to normalization).

Back to isotropic covariance functions

- Isotropic covariance functions, i.e., $C(\|\mathbf{h}\|)$, where $\|\mathbf{h}\|$ is length of $\mathbf{h}$, are the most frequent choice within stationarity.
- Several discussed in the first lecture
- An isotropic covariance function that is valid in dimension r need not be valid in dimension $r+1$.
- Intuition by considering $r = 1$ versus $r = 2$. For three points, in one-dimensional space, given the distances separating points 1 and 2 (d_{12}) and points 2 and 3 (d_{23}), then the distance separating points 1 and 3 d_{13} is either $d_{12} + d_{23}$ or $|d_{12} - d_{23}|$.
- But in two-dimensional space, given d_{12} and d_{23}, d_{13} can take any value in $\Re^+$ (subject to the triangle inequality)

Back to isotropic covariance functions

- There are isotropic correlation functions that are valid in all dimensions. The Gaussian correlation function, $\rho(\|h\|) = \exp(-\phi \|h\|^2)$ is an example.
- It is the characteristic function associated with r i.i.d. normal random variables, each with variance $1/(2\phi)$ for any r.
- More generally, the powered exponential, $\exp(-\phi \|h\|^\alpha)$, $0 < \alpha \leq 2$ (and hence the exponential correlation function) is valid for any r
- A general result is that $C(||\mathbf{h}||)$ is a positive definite isotropic function on $\Re^r$ for all r if and only if it has the representation, $C(||\mathbf{h}||) = \int e^{-w||\mathbf{h}||^2} G(dw)$ where G is nondecreasing and bounded and $w \in R^+$.
- So, $C(||\mathbf{h}||)$ arises as a scale mixture of Gaussian correlation functions.
- G might be a c.d.f. on R^+ with a p.d.f., $g(w)$, i.e., $G(dw) = g(w)dw$.

Back to isotropic covariance functions

- All valid isotropic correlation functions in dimension r.
- From Matérn, the set of $C(\|h\|)$ is of the form

$$C(\|h\|) = \int_0^\infty \left(\frac{2}{w\,\|\mathbf{h}\|}\right)^\alpha \Gamma(\nu+1) J_\nu(w\,\|\mathbf{h}\|) G(dw)\,,$$

  where G is nondecreasing and integrable on $\Re^+$, J_ν is the Bessel function of the first kind of order ν, and $\nu = (r-2)/2$ provides all valid isotropic correlation functions on $\Re^r$.
- When $r = 2$, $v = 0$; arbitrary correlation functions in two-dimensional space arise as scale mixtures of Bessel functions of order 0.
- $J_0(d) = \sum_{k=0}^\infty \frac{(-1)^k}{(k!)^2}\left(\frac{d}{2}\right)^{k/2}$. J_0 decreases from 1 at $d = 0$ and will oscillate above and below 0 with amplitudes and frequencies that are diminishing as d increases
- Typically, correlation functions that are monotonic and decreasing to 0 are chosen but, apparently, valid correlation functions can permit negative associations.

The range

- If we confine ourselves to strictly monotonic isotropic covariance functions can formalize the notion of a range.
- The range is conceptualized as the distance beyond which association becomes negligible.
- If the covariance function reaches 0 in a finite distance, then we refer to this distance as the range.
- However, we customarily work with covariance functions that attain 0 asymptotically as $\|\mathbf{h}\| \to \infty$.
- In this case, it is common to define the range as the distance beyond which correlation is less than .05.
- So if ρ is the correlation function, then writing the range as R we solve $\rho(R; \boldsymbol{\theta}) = .05$, where $\boldsymbol{\theta}$ denotes the parameters in the correlation function.
- Therefore, R is an implicit function of the parameter $\boldsymbol{\theta}$.

Smoothness of process realizations

- How does one select among the various choices of correlation functions? Usual model selection criteria will typically find it difficult to distinguish, say, among one-parameter isotropic scale choices such as the exponential, Gaussian, or Cauchy.
- Through suitable alignment of parameters, the correlation curves will be very close to each other.
- An alternative perspective is to make the selection based upon theoretical considerations. This arises from the fact that the choice of correlation function determines the smoothness of realizations from the spatial process.
- More precisely, a process realization is viewed as a random surface over the region. By choice of C we can ensure that these realizations will be almost surely continuous, or mean square continuous, or mean square differentiable, and so on.

Smoothness of process realizations

- However, at best the process is only observed at finitely many locations. (At worst, it is never observed, e.g., when the spatial process is used as a second stage model for random spatial effects.) So, it is not possible to "see" the smoothness of the process realization.
- Elegant theory, developed in Kent, Stein, and extended in Banerjee and Gelfand clarifies the relationship between the choice of correlation function and such smoothness.

Smoothness of process realizations

- The key point is that, according to the process being modeled, we may anticipate surfaces to not be continuous (as with digital elevation models in the presence of gorges, escarpments, or other topographic features), or to be differentiable (as in studying land value gradients or temperature gradients).
- Can choose a correlation function to ensure such behavior.
- Recall the Matérn class; ν is a smoothness parameter.
- In two dimensions, greatest integer in ν indicates the number of times process realizations will be mean square differentiable.
- Use of the Matérn covariance function as a model enables the data to inform about ν; we can learn about process smoothness despite observing the process at only a finite number of locations.
- Hence, we recommend the Matérn class as a general isotropic specification for building spatial models.

Smoothness of process realizations

- Through the inversion formula

$$2\left(\frac{\phi\,\|\mathbf{h}\|}{2}\right)^{\nu}\frac{K_\nu(\phi(\|\mathbf{h}\|))}{\phi^{2\nu}\Gamma(\nu+\frac{r}{2})}=\int_{\Re^r}e^{i\mathbf{w}^T\mathbf{h}}(\phi^2+\|\mathbf{w}\|^2)^{-(\nu+r/2)}d\mathbf{w}\,,$$

  where K_ν is the modified Bessel function of order $\nu > 0$.
- The Matérn covariance function arises as the characteristic function from a Cauchy spectral density.
- We return to the question of how rich is the class of stationary processes obtained using kernel mixing?
- Elegant result (Yaglom, 1987): A stationary random process can be defined by kernel mixing iff it has a spectral density.
- With the Matérn covariance function we can show that only $\nu > 1$ can arise from kernel convolution; we can not create the exponential covariance from kernel mixing.

Directional derivative processes

- So, we have the connection between the correlation function and the smoothness of process realizations.
- When realizations are mean square differentiable, we can think about a directional derivative process.
- At each location we can define a random variable that is the directional derivative of the original process at that location in the given direction. The entire collection of random variables is a spatial process.
- Intuitively, such variables would be created through limits of finite differences, i.e., we can also formalize a finite difference process in a given direction.
- A directional derivative process enables assessing where there are sharp gradients and in which directions.
- Applications: land-value gradients away from a central business district, temperature gradients in a north-south direction, maximum gradient at a location and associated direction for zones of rapid change (boundary analysis).

Spatial gradient analysis

- First, a review of basic gradient ideas
- Gradients associated with Gaussian process realizations (focus on spatial processes, i.e., R^2)
- Smoothness suggests working with the Matérn correlation function
- "Large scale" inference for the surface - prediction, global behavior, where high, where low
- "Small scale" behavior - local, slopes, gradients, boundaries

Directional Gradients

- "Finite differences" for any direction (unit vector) $\mathbf{u}$

$$Y_{\mathbf{u},h}(\mathbf{s}) = \frac{Y(\mathbf{s}+h\mathbf{u}) - Y(\mathbf{s})}{h},$$

  This is a process
- Pass to "limit" to consider "local slope" process

$$D_{\mathbf{u}}Y(\mathbf{s}) = \lim_{h\to 0} Y_{\mathbf{u},h}(\mathbf{s}) = \lim_{h\to 0} \frac{Y(\mathbf{s}+h\mathbf{u}) - Y(\mathbf{s})}{h}.$$

Mean square differentiability

- At location $\mathbf{s}$ need a vector $\nabla Y(\mathbf{s})$, such that for every unit vector $\mathbf{u}$ and scalar h

$$Y(\mathbf{s}+h\mathbf{u}) = Y(\mathbf{s}) + h\mathbf{u}^T \nabla Y(\mathbf{s}) + R(\mathbf{s}, h\mathbf{u}), \;\; R(\mathbf{s}, h\mathbf{u}) \to 0$$

  as $h \to 0$ (in L^2 sense).
- Then $D_{\mathbf{u}} Y(\mathbf{s}) = \mathbf{u}^T \nabla Y(\mathbf{s})$
- Under stationarity: existence of $H_C = \left(\left(\frac{\partial^2 C}{\partial s_i \partial s_j} \right) \right)$ at $\mathbf{0} \Longrightarrow$ existence of $D_{\mathbf{u}} Y(\mathbf{s})$.
- Smoothness of covariance function $\Longrightarrow$ smoothness of process realizations

Basis directions

- $\{\mathbf{e}_1, .., \mathbf{e}_d\}$ is an o.n basis for R^d
- $\mathbf{u} = \sum w_i \mathbf{e}_i$; $w_i = \mathbf{u}^T \mathbf{e}_i$.
-

$$D_{\mathbf{u}} Y(\mathbf{s}) = \mathbf{u}^T \nabla_Y(\mathbf{s}) = \sum_{i=1}^{d} w_i \mathbf{e}_i^T \nabla_Y(\mathbf{s}) = \sum w_i D_{\mathbf{e}_i} Y(\mathbf{s})$$

- So $\{D_{\mathbf{e}_i} Y(\mathbf{s})\}$ is a basis (of random functions) and $\nabla_Y(\mathbf{s}) = \nabla Y(\mathbf{s})$ - gradient vector.

- Only need a basis set of directions.
- Maximum gradient value: $||\nabla Y(\mathbf{s})||$
- Direction of maximum gradient: $\nabla Y(\mathbf{s}) / ||\nabla Y(\mathbf{s})||$

Distribution theory

- Want to predict in many directions, say $D_{\mathbf{u}_1} Y(\mathbf{s}_0), ..., D_{\mathbf{u}_p} Y(\mathbf{s}_0)$
- But only need to predict $\nabla Y(\mathbf{s}) = (D_{\mathbf{e}_1} Y(\mathbf{s}_0), ..., D_{\mathbf{e}_d} Y(\mathbf{s}_0))$.
- $Y(\mathbf{s})$: (0 mean, cov. func. C) Gaussian process admits derivative process
- Then $D_{\mathbf{u}} Y(\mathbf{s})$ is Gaussian (0 mean) process, covariance function:

$$C_{\mathbf{u}}(\mathbf{s}, \mathbf{s}') = -\mathbf{u}^T H_C(\mathbf{\Delta})\,\mathbf{u}, \text{ where } \mathbf{\Delta} = \mathbf{s} - \mathbf{s}';$$
$$(H_C(\mathbf{\Delta}))_{ij} = \partial^2 C(\mathbf{\Delta}) / \partial\Delta_i \partial\Delta_j.$$

Nonstationary spatial process models

- Choices that offer attractive interpretation and are computationally tractable. So, we prefer constructive approaches.
- Nonstationarity can be immediately introduced through scaling and through marginalization of stationary processes.
- Suppose $w(\mathbf{s})$ is a mean 0, variance 1 stationary process with correlation function ρ. Then $v(\mathbf{s}) = \sigma(\mathbf{s})w(\mathbf{s})$ is a nonstationary process. In fact,

$$\begin{aligned} var\ v(\mathbf{s}) &= \sigma^2(\mathbf{s}) \\ \text{and } cov(v(\mathbf{s}), v(\mathbf{s}')) &= \sigma(\mathbf{s})\sigma(\mathbf{s}')\rho(\mathbf{s}-\mathbf{s}')\ , \end{aligned}$$

- Set $\sigma(\mathbf{s}) = g(X(\mathbf{s}))\sigma$ where $X(\mathbf{s})$ is a suitable positive covariate and g is a strictly increasing positive function.

Nonstationary spatial process models

- Suggests a simple strategy for developing nonstationary covariance structure using known functions. For instance, for a function $g(\mathbf{s})$ on $\Re^2$, $C(\mathbf{s}, \mathbf{s}') = \sigma^2 g(\mathbf{s}) g(\mathbf{s}')$ is immediately seen to be a valid covariance function.
- Not very interesting since, for locations $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n$, the resulting joint covariance matrix is of rank 1 regardless of n!
- But leads to the flexible class of nonstationary covariance functions introduced by Cressie and Johannesson (2008) to implement so-called fixed rank kriging
- Let $C(\mathbf{s}, \mathbf{s}') = \mathbf{g}(\mathbf{s})^T K \mathbf{g}(\mathbf{s}')$ where $\mathbf{g}(\mathbf{s})$ is an $r \times 1$ vector of known functions and K is an $r \times r$ positive definite matrix. Again, the validity of this C is immediate.
- No requirement that the functions in $\mathbf{g}(\mathbf{s})$ be orthogonal and standard classes such as smoothing splines, radial basis functions, or wavelets can be used.
- The challenges include the choice of r and the estimation of K (no Bayesian version available)

Deformation

- An approach to nonstationarity through *deformation* (Sampson and Guttorp).
- Transform the geographic region D to a new region G, a region such that stationarity and, in fact, isotropy holds on G. The mapping $\mathbf{g}$ from D to G is bivariate, i.e., if $\mathbf{s} = (\ell_1, \ell_2)$, $\mathbf{g}(\ell_1, \ell_2) = (g_1(\ell_1, \ell_2), g_2(\ell_1, \ell_2))$. If C denotes the isotropic covariance function on G we have

$$cov(Y(\mathbf{s}), Y(\mathbf{s}')) = C(\|\mathbf{g}(\mathbf{s}) - \mathbf{g}(\mathbf{s}')\|) .$$

- There are two unknown functions to estimate, $\mathbf{g}$ and C. The latter is assumed to be a parametric choice from a standard class of covariance functions.
- To determine the former is a challenging "fitting" problem. To what class of transformations shall we restrict ourselves?
- Sampson and Guttorp optimize over thin plate splines
- Lots of follow on work: Smith; Damian, Sampson, Guttorp; Schmidt and O'Hagan

Nonstationary spatial process models

- A fundamental limitation of the deformation approach is that implementation requires independent replications of the process in order to obtain an estimated sample covariance matrix for the set of $(Y(\mathbf{s}), ..., Y(\mathbf{s}_n))$.
- If we obtain repeated measurements at a particular location, they are typically collected across time. We would prefer to incorporate a temporal aspect in the modeling
- Moreover, even if we assume independent replications, to estimate well an $n \times n$ covariance matrix, even for a moderate size n, requires a very large number of them, more than we would imagine in practice.

Nonstationarity through kernel mixing of process variables

- Recall that kernel mixing was described above in the context of creating stationary processes.
- A strategy for introducing nonstationarity while retaining clear interpretation and permitting analytic calculation.
- Kernel mixing is often done with distributions and we will look at this idea as well.
- First, consider stationary choices of the form $K(\mathbf{s}-\mathbf{s}')$, e.g., $K(\mathbf{s}-\mathbf{s}') = \exp\{-\frac{1}{2}(\mathbf{s}-\mathbf{s}')^T V(\mathbf{s}-\mathbf{s}')\}$.
- Choice for V would be diagonal with V_{11} and V_{22} providing componentwise scaling to the separation vector $\mathbf{s}-\mathbf{s}'$.
- Let $Z(\mathbf{s})$ be a white noise process, i.e., $E(Z(\mathbf{s}) = 0)$, $var(Z(\mathbf{s})) = \sigma^2$ and $cov(Z(\mathbf{s}), Z(\mathbf{s}')) = 0$ and set

$$w(\mathbf{s}) = \int_{\Re^2} K(\mathbf{s}-\mathbf{t})Z(\mathbf{t})d\mathbf{t}\ .$$

- Rigorously, the convolution should be written as $w(\mathbf{s}) = \int K(\mathbf{s}-\mathbf{t})\mathcal{X}(d\mathbf{t})$ where $\mathcal{X}(\mathbf{t})$ is two-dimensional Brownian motion.

Kernel mixing

- We have $E[w(\mathbf{s})] = 0$, but also

$$var\ w(\mathbf{s}) = \sigma^2 \int_{\Re^2} K^2(\mathbf{s}-\mathbf{t})d\mathbf{t}\ ,$$
$$\text{and} \quad cov(w(\mathbf{s}), w(\mathbf{s}')) = \sigma^2 \int_{\Re^2} K(\mathbf{s}-\mathbf{t})K(\mathbf{s}'-\mathbf{t})d\mathbf{t}\ .$$

- A change of variables ($\mathbf{t} \to \mathbf{u} = \mathbf{s}' - \mathbf{t}$), shows that

$$cov(w(\mathbf{s}), w(\mathbf{s}')) = \sigma^2 \int_{\Re^2} K(\mathbf{s}-\mathbf{s}'+\mathbf{u})K(\mathbf{u})d\mathbf{u}\ ,$$

  i.e., $w(\mathbf{s})$ is stationary.
- More generally, if $z(\mathbf{s})$ is a mean 0 stationary spatial process with covariance function $\sigma^2\rho(\cdot)$ then, again $E[w(\mathbf{s})] = 0$ but

$$var\ w(\mathbf{s}) = \sigma^2 \int_{\Re^2}\int_{\Re^2} K(\mathbf{s}-\mathbf{t})K(\mathbf{s}'-\mathbf{t})\rho(\mathbf{t}-\mathbf{t}')d\mathbf{t}d\mathbf{t}'$$
$$cov(w(\mathbf{s}), w(\mathbf{s}')) = \sigma^2 \int_{\Re^2}\int_{\Re^2} K(\mathbf{s}-\mathbf{t})K(\mathbf{s}'-\mathbf{t}')\rho(\mathbf{t}-\mathbf{t}')d\mathbf{t}d\mathbf{t}'\ .$$

  $w(\mathbf{s})$ is still stationary.
- Above can be can be proposed as covariance functions. However, explicit integrations will not be possible except in certain special cases

Kernel mixing

- An alternative is a finite sum approximation,

$$w(\mathbf{s}) = \sum_{j=1}^{L} K(\mathbf{s} - \mathbf{t}_j) Z(\mathbf{t}_j)$$

  for locations $\mathbf{t}_j$, $j = 1, \ldots, L$
- We have

$$var\ w(\mathbf{s}) = \sigma^2 \sum_{j=1}^{L} \sum_{j'=1}^{L} K(\mathbf{s} - \mathbf{t}_j) K(\mathbf{s} - \mathbf{t}_{j'}) \rho(\mathbf{t}_j - \mathbf{t}_{j'})$$
$$cov(w(\mathbf{s}), w(\mathbf{s}')) = \sigma^2 \sum_{j=1}^{L} \sum_{j'=1}^{L} K(\mathbf{s} - \mathbf{t}_j) K(\mathbf{s}' - \mathbf{t}_{j'}) \rho(\mathbf{t}_j - \mathbf{t}_{j'})$$

- Finite sum process is no longer stationary but artificial as it arises from the arbitrary $\{\mathbf{t}_j\}$.
- Instead, suppose we allow the kernel to vary spatially, i.e., $K(\mathbf{s} - \mathbf{s}'; \mathbf{s})$.

Kernel mixing

- We might take $K(\mathbf{s}-\mathbf{s}';\mathbf{s}) = \exp\{-\frac{1}{2}(\mathbf{s}-\mathbf{s}')^T V_{\mathbf{s}}(\mathbf{s}-\mathbf{s}')\}$. As above, we might take $V_{\mathbf{s}}$ to be diagonal with, if $\mathbf{s} = (\ell_1, \ell_2)$, $(V_{\mathbf{s}})_{11} = V(\ell_1)$, and $(V_{\mathbf{s}})_{22} = V(\ell_2)$.
- Now, the process is nonstationary.
- A nice extension (Paciorek and Schervish) who note that the general form $C(\mathbf{s},\mathbf{s}') = \int_{\Re^2} K_{\mathbf{s}}(\mathbf{u})K_{\mathbf{s}'}(\mathbf{u})d\mathbf{u}$ is a valid covariance function.
- Avoiding Gaussian kernels, let $Q(\mathbf{s},\mathbf{s}') = (\mathbf{s}-\mathbf{s}')^T(\frac{V_{\mathbf{s}}+V_{\mathbf{s}'}}{2})^{-1}(\mathbf{s}-\mathbf{s}')$ and let ρ be any positive definite function on R^2. Then

$$C(\mathbf{s},\mathbf{s}') = |V_{\mathbf{s}}|^{\frac{1}{2}}|V_{bs'}|^{\frac{1}{2}}|\frac{V_{\mathbf{s}}+V_{\mathbf{s}'}}{2}|^{-\frac{1}{2}}\rho(\sqrt{Q(\mathbf{s},\mathbf{s}')})$$

is a valid nonstationary correlation function.

Mixing of process distributions

- Suppose a kernel $K(\cdot)$ is integrable and standardized to a density and f is also a density function, then

$$f_K(y) = \int K(y-x)f(x)dx$$

  is a density function.
- It is the distribution of $X + Y - X$ where $X \sim f$, $Y - X \sim k$, and X and $Y - X$ are independent.
- To extend to arbitrary finite dimensional distributions, let $\mathbf{Y}$, a vector of dimension n and use above to build a process distribution.
- Operating formally, let V_D be the set of all $V(\mathbf{s})$, $\mathbf{s} \in D$. Write $V_D = V_{0,D} + V_D - V_{0,D}$ where $V_{0,D}$ is a realization of a mean 0 stationary Gaussian process over D, and $V_D - V_{0,D}$ is a realization of a white noise process with variance τ^2 over D. Write

$$f_K(V_D \mid \tau) = \int \frac{1}{\tau} K\left(\frac{1}{\tau}(V_D - V_{0,D})\right) f(V_{0,D})dV_{0,D}$$

Kernel mixing

- Formally, f_K is the distribution of the spatial process $V(\mathbf{s})$.
- In fact, $V(\mathbf{s})$ is just the customary model for the residuals in a spatial regression, i.e., of the collection $V(\mathbf{s}) = w(\mathbf{s}) + \epsilon(\mathbf{s})$ where $w(\mathbf{s})$ is a spatial process and $\epsilon(\mathbf{s})$ is a noise or nugget process
- Reveals how a spatial process can be developed through "kernel mixing" of a process distribution using an alternative specification for $V_{0,D}$.
- If $f(V_{0,D})$ is a discrete distribution, say, of the form $\sum_\ell p_\ell \delta(V^*_{\ell,D})$ where $p_\ell \geq 0$, $\sum p_\ell = 1$, $\delta(\cdot)$ is the Dirac delta function, and $V^*_{\ell,D}$ is a surface over D.
- We have a nonstationary process with easy distribution theory and moment calculations