Beyond Matérn: on the class of confluent hypergeometric covariance functions for Gaussian process modeling

Anindya Bhadra
www.stat.purdue.edu/∼bhadra

Purdue University

Overview

- Mean zero Gaussian processes (GP) are completely determined by their covariance functions.
- There is an over-reliance in spatial statistics and GP literature in general on the Matérn covariance function. There are good reasons for this. We first discuss the strengths and limitations of Matérn.
- We then propose a new interpretable covariance function keeping the strengths of Matérn and rectifying one major limitation.
- Preprint at: `https://arxiv.org/abs/1911.05865`. Joint work with Pulong Ma (SAMSI & Duke University). Supported by NSF Grant DMS-2014371.

Gaussian processes and covariance functions

- A stochastic process $\{Z(\mathbf{s}) \in \mathbb{R} : \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d\}$ is a Gaussian process (GP) if every finite-dimensional realization $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$ jointly follows a multivariate normal distribution for $\mathbf{s}_i \in \mathcal{D}$ and every n.

- The properties of a GP are determined completely by the mean and covariance functions. In this talk, the mean is assumed zero throughout.

- $Z(\cdot)$ is said to be a *stationary* process with a covariance function $\mathrm{cov}(Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})) = C(\mathbf{h})$ if $C(\cdot)$ solely a function of the increment $\mathbf{h}$.

- If $C(\cdot)$ is a function of $\|\mathbf{h}\|$ with $\|\cdot\|$ denoting the Euclidean norm, then $C(\cdot)$ is called isotropic.

Origin and tail behavior of covariance functions

- The behavior of the covariance function at small distances determines the smoothness behavior of the process (formally, the degree of the mean squared differentiability).
- The behavior of the covariance function at large distances determines whether distant observations are allowed to be correlated (long range dependence).
- Many popular covariance functions (e.g., squared exponential) are inflexible, in the sense that processes using these covariances are infinitely mean squared differentiable (very smooth).
- Computer experiments researchers are usually OK with that but Stein (1999) thinks very smooth spatial processes are unrealistic.

Matérn

- The isotropic Matérn covariance function is of the form

$$\mathcal{M}(h) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\phi} h \right)^{\nu} \mathcal{K}_{\nu} \left(\frac{\sqrt{2\nu}}{\phi} h \right),$$

  where $\sigma^2 > 0$ is the variance parameter, $\phi > 0$ is the range parameter, and $\nu > 0$ is the smoothness parameter that controls the differentiability of the associated random process.

- Processes with a Matérn covariance function are exactly $\lfloor \nu \rfloor$ times differentiable in the mean squared sense.

- This precise control over smoothness via ν is a key reason for the popularity of Matérn, as is the interpretability of other parameters.

- Stein (1999) makes a summary recommendation to conclude his first chapter: "Use the Matérn model."

Tail decay of Matérn

- For large h, the behavior of the Matérn covariance function is given by:

$$\mathcal{M}(h) \asymp h^{\nu-1/2} \exp\left(-\frac{\sqrt{2\nu}}{\phi} h\right), \quad h \to \infty.$$

- Eventually, the $\exp(-\sqrt{2\nu}h/\phi)$ term dominates, and the covariance decays exponentially for large h, no flexibility there!

- May be inappropriate for settings where processes display high correlation at large distances.

Polynomial-tailed covariances, e.g., Cauchy

- If exponential tail decay is a problem, of course one may use a polynomial-tailed covariance function (e.g., generalized Cauchy).
- However, the cost of switching to polynomial covariances is great!
- Processes using a Cauchy covariance function are either infinitely mean squared differentiable (very smooth) or not at all (very rough). There is no middle ground. Same issue with power exponential family.
- Our proposal: a new covariance function with interpretable parameters that has the same origin behavior as Matérn and hence, allows precise control over smoothness; but also has polynomial tails.

Motivating example: the OCO-2 data

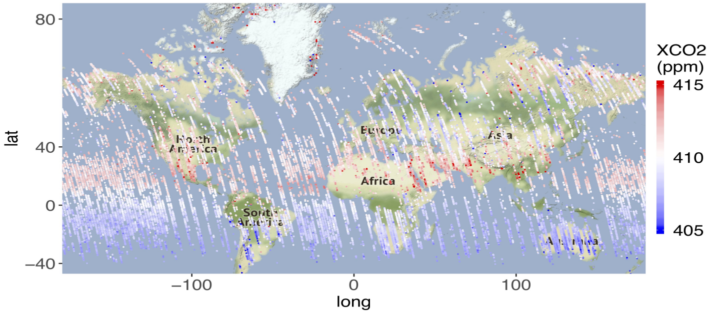


Figure: XCO2 data from June 1 to June 16, 2019. The units are ppm.

- Spatial process: unlikely to be very smooth.
- Large gaps between longitude bands. If a rapidly decaying covariance function is used, harder to borrow information across large distances.

Matérn + Cauchy?

- Porcu and Stein (2012) propose a simple solution: use a summed Matérn and Cauchy covariance, which is a valid covariance function!
- Matérn dominates near the origin, Cauchy dominates at the tails.
- We identify at least three problems with this:
  - Parameter interpretability is lost and there are more parameters than needed: e.g., one scale parameter from Matérn and one from Cauchy.
  - Consequently, numerical optimization is hard.
  - The microergodic parameter (i.e., the parameter that is consistently estimable under infill asymptotics) has no closed form.
- The trouble is, finite mixtures like this are in general cumbersome. Infinite mixtures are often more pleasant.

Mixture of Matérn: construction (and some intuition)

- *Our key innovation* is to note the correspondence between the GIG normalizing constant of Barndorff-Nielsen (1977) and Matérn:

$$\pi_{GIG}(x) = \frac{(a/b)^{p/2}}{2\mathcal{K}_p(\sqrt{ab})} x^{(p-1)} \exp\{-(ax + b/x)/2\}; \quad a, b > 0, p \in \mathbb{R}.$$

Thus,

$$\mathcal{K}_p(\sqrt{ab}) = \frac{1}{2}(a/b)^{p/2} \int_0^\infty x^{(p-1)} \exp\{-(ax + b/x)/2\} dx.$$

- Take $a = \phi^{-2}, b = 2\nu h^2$ and $p = \nu$. This yields:

$$\begin{aligned} \mathcal{M}(h) &= \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\phi} h\right)^\nu \mathcal{K}_\nu \left(\frac{\sqrt{2\nu}}{\phi} h\right) \\ &= \frac{\sigma^2}{2^\nu \phi^{2\nu} \Gamma(\nu)} \int_0^\infty x^{(\nu-1)} \exp\{-(x/\phi^2 + 2\nu h^2/x)/2\} dx. \end{aligned}$$

Mixture of Matérn: construction (and some intuition)

- With this GIG integral representation of Matérn, our construction is simply a mixture over ϕ^2:

$$
\begin{aligned}
C(h) &= \int_0^\infty \mathcal{M}(h) dG(\phi^2) \\
&= \int_0^\infty \left[\frac{\sigma^2}{2^\nu \phi^{2\nu} \Gamma(\nu)} \int_0^\infty x^{(\nu-1)} \exp\{-(x/\phi^2 + 2\nu h^2/x)/2\} dx \right] dG(\phi^2) \\
&\overset{\text{Fubini}}{=} \frac{\sigma^2}{2^\nu \Gamma(\nu)} \int_0^\infty x^{(\nu-1)} \left[\int_0^\infty \phi^{-2\nu} \exp\{-x/(2\phi^2)\} dG(\phi^2) \right] \exp(-\nu h^2/x) dx.
\end{aligned}
$$

- Main intuition: the outer integral is a Gaussian variance mixture with respect to x. A result of Barndorff-Nielsen et al. (1982) allows us to connect the tail behavior of $C(h)$ with that of the mixing density for x, which in turn depends on the inner integral over ϕ^2.

- Recall, we want polynomial tail for $C(h)$. Barndorff-Nielsen allows us to simply *reverse engineer* the $dG(\phi^2)$ mixing density to achieve this!

Mixture of Matérn: construction (and some intuition)

- Following these tail considerations, we arrive at: $\phi^2 \sim \mathcal{IG}(\alpha, \beta/2)$. This is far from accidental.

- Fortuitously, this mixing leaves the flexible origin behavior of Matérn intact, otherwise this would be no better than (generalized) Cauchy!

- Finally, we are ready to present that: $C(h) = \int_0^\infty \mathcal{M}(h)\ \pi(\phi^2) d\phi^2$ is a valid covariance function on $\mathbb{R}^d$ with the following form:

$$C(h) = \frac{\sigma^2 \beta^\alpha \Gamma(\nu+\alpha)}{\Gamma(\nu)\Gamma(\alpha)} \int_0^\infty x^{(\nu-1)}(x+\beta)^{-(\nu+\alpha)} \exp\left(-\nu h^2/x\right) dx,$$

  where $\sigma^2 > 0$ is the variance parameter, $\alpha > 0$ is the tail decay parameter, $\sqrt{\beta} > 0$ is the range parameter, and $\nu > 0$ is the smoothness parameter.

- One more parameter (α) compared to Matérn that controls the polynomial tail decay.

Mixture of Matérn: the confluent hypergeometric class

- The integral on the last slide looks daunting, but there is an equivalent representation in terms of a special function that can be evaluated in other ways, e.g., via a quickly converging infinite series.

- In particular, it turns out we have

$$C(h) = \frac{\sigma^2 \Gamma(\nu + \alpha)}{\Gamma(\nu)} \mathcal{U}(\alpha, 1 - \nu, \nu h^2 / \beta),$$

  where $\mathcal{U}(a, b, c)$ is the confluent hypergeometric function of the second kind.

- Hence, we call the new covariance class the confluent hypergeometric (CH) class.

- Efficient implementations are available in R, MATLAB or GSL.

The CH class: parameter interpretability

- Arguably the key distinguishing feature from the summed Matérn + Cauchy covariance is that we have 4 parameters, each with a specific interpretation.

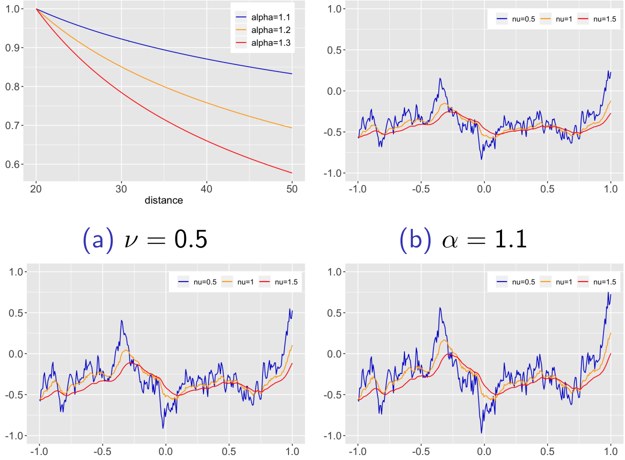
- Easy to check that $C(0) = \sigma^2$, i.e., σ^2 is the variance parameter.

- Similarly, $\sqrt{\beta}$ is the range parameter, since $C(\cdot)$ depends of h^2/β.

- But MOST IMPORTANTLY, we establish the following two properties:
  (a) **Origin behavior**: The differentiability of the CH class is solely controlled by ν in the same way as the Matérn class.
  (b) **Tail behavior**: $C(h) \asymp |h|^{-2\alpha} L(h^2)$ as $h \to \infty$, where $L(\cdot)$ is a slowly varying function at ∞.

- Origin behavior depends on ν, tail behavior depends on α, independently of each other.

Process realizations



(a) $\nu = 0.5$ (b) $\alpha = 1.1$

(c) $\alpha = 1.2$ (d) $\alpha = 1.3$

Figure: Realizations from a Gaussian process with zero mean and the proposed covariance function on a 1-dimensional domain. In panel (a), the y-axis is the quantity $C(h)/C(20)$ since this cancels out the constants.

Properties of the CH class I: tail behavior of the spectral density

- The spectral density of the CH covariance function admits the following tail behavior:

$$f(\omega) \sim \frac{\sigma^2 2^{2\nu} \nu^\nu \Gamma(\nu+\alpha)}{\pi^{d/2} \beta^\nu \Gamma(\alpha)} \omega^{-(2\nu+d)} L(\omega^2), \quad \omega \to \infty,$$

  where $L(x) = \{x/(x+\beta/(2\nu))\}^{\nu+d/2}$ is slowly varying at ∞.

- For large ω, this is the spectral density of Matérn (given by $\omega^{-(2\nu+d)}$) multiplied by a slow function, up to a constant.

- Makes sense because tail behavior of the spectral density corresponds to the origin behavior of the covariance function, and the origin behaviors of the CH and Matérn classes are similar.

Properties of the CH class II: the microergodic parameter

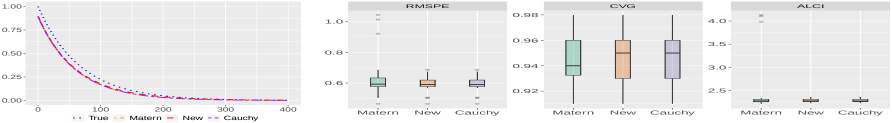
- Under infill asymptotics, the individual parameters in a covariance function are not typically consistently estimable, instead some combination of them is.
- Zhang (2004, JASA) derives this in closed form for Matérn. It is: $\sigma^2\phi^{-2\nu}$.
- We are also able to derive it in closed form for the CH class. It is:

$$\frac{\sigma^2\beta^{-\nu}\Gamma(\nu+\alpha)}{\Gamma(\alpha)}.$$

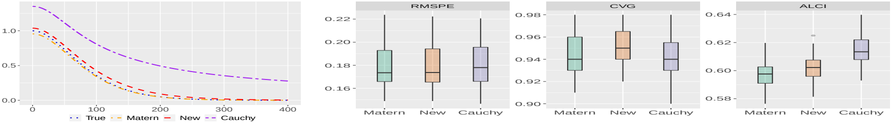
Properties of the CH class III: asymptotic normality of the mle

- We are able to establish asymptotic normality results for the maximum likelihood estimates of the parameters.
- There are also additional results on the asymptotic efficiency of the kriging estimates and on equivalence of Gaussian measures under the CH and Matérn classes.
- See paper for details.

Case I: True covariance is Matérn



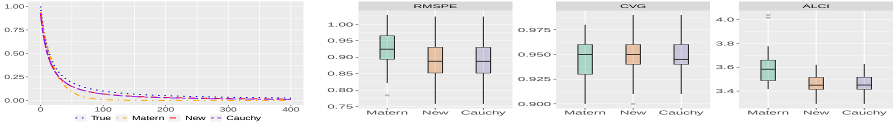
$\nu = 0.5$, $ER = 200$



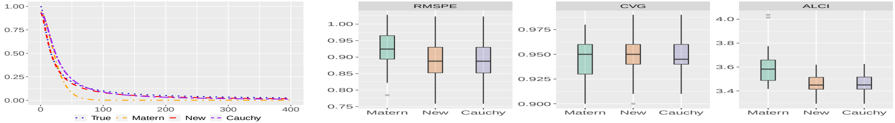
$\nu = 2.5$, $ER = 200$

Figure: Case 1: Comparison of predictive performance and estimated covariance structures when the true covariance is the Matérn class with 2000 observations. The predictive performance is evaluated at 10-by-10 regular grids in the square domain. These figures summarize the predictive measures based on RMSPE, CVG and ALCI under 30 simulated realizations.

Case II: True covariance is CH



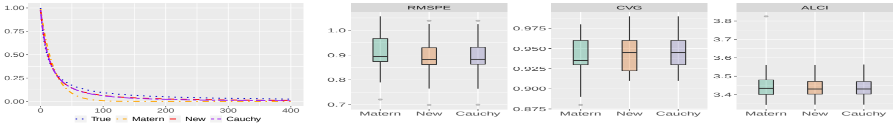
$\nu = 0.5$, $ER = 200$



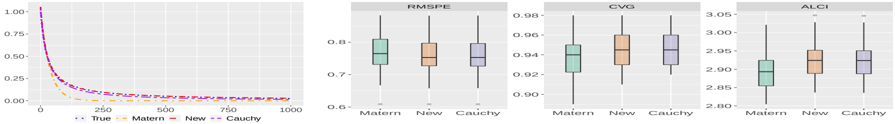
$\nu = 2.5$, $ER = 200$

Figure: Case 2: Comparison of predictive performance and estimated covariance structures when the true covariance is the CH class with 2000 observations. The predictive performance is evaluated at 10-by-10 regular grids in the square domain. These figures summarize the predictive measures based on RMSPE, CVG and ALCI under 30 simulated realizations.

Case III: True covariance is Cauchy



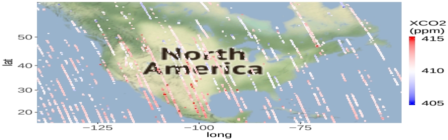
$\delta = 1, ER = 200$



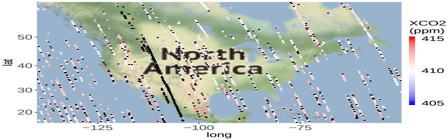
$\delta = 1, ER = 500$

Figure: Case 3: Comparison of predictive performance and estimated covariance structures when the true covariance is the GC class with 2000 observations. The predictive performance is evaluated at 10-by-10 regular grids in the square domain. These figures summarize the predictive measures based on RMSPE, CVG and ALCI under 30 simulated realizations.

Analysis of the OCO-2 data



(a) XCO2 data in the study region.



(b) XCO2 testing data in black.

Figure: XCO2 measurements from June 1 to June 16, 2019 in the study region.

- Given the data $\mathbf{Z} := (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^\top$, we assume a typical spatial process model:

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s}), \quad \mathbf{s} \in \mathcal{D},$$

  where $Y(\cdot)$ is a GP with mean function $\mu(\cdot)$ and covariance $C(\cdot, \cdot)$.

- The term $\epsilon(\cdot)$ is assumed to be a spatial white noise process accounting for the nugget effect with $\text{var}(\epsilon(\mathbf{s})) = \tau^2 > 0$.

Analysis of the OCO-2 data

Table: Cross-validation results on the XCO2 data based on the Matérn and CH covariance models. The numbers in the first coordinate correspond to those based on MAR locations for interpolative prediction, and the numbers in the second coordinate correspond to those based on MBD locations for extrapolative prediction.

	Matérn class		CH class	
	$\nu = 0.5$	$\nu = 1.5$	$\nu = 0.5$	$\nu = 1.5$
τ^2 (nugget)	0.0642	0.2215	**0.0038**	0.1478
RMSPE	0.672, 1.478	0.675, 1.599	0.676, 1.263	0.735, 1.227
CVG(95%)	0.952, 0.929	0.952, 0.951	0.944, 0.921	0.878, 0.937
ALCI(95%)	2.533, 5.095	2.536, 5.044	2.543, 4.722	2.098, 4.855

Additional results, summary and future directions

- We extensively investigated tensor products and other derived covariances from the CH class; the results are very encouraging (see paper).
- This means, the proposed covariance class may be an important tool in the study of computer experiments with larger dimensions, although we only studied a spatial (2-d) application.
- In summary: the CH covariance class is more flexible in modeling distant correlations than Matérn, without sacrificing the control over smoothness.
- Applications to computer experiments and extensions to space-time covariance functions appear promising future directions.

Main references

- Ma, P. and **Bhadra, A.** (2020+). Kriging: Beyond Matérn. (under revision). *arXiv preprint* arXiv:1911.05865.
- Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* Springer, New York, NY.
- Barndorff-Nielsen, O. E. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Royal Society of London Proceedings Series A* **353**, 401–419.
- Barndorff-Nielsen, O. E., Kent, J. T., and Sørensen, M. (1982). Normal variance-mean mixtures and z distributions. *International Statistical Review* **50**, 145–159.