Kernels for deterministic and stochastic approximations of (invariant) functions

David Ginsbourger

[1] Idiap Research Institute, UQOD group, Martigny, Switzerland, and
[2] IMSV, Mathematics and Statistics Department, University of Bern, Switzerland

Acknowledgements: a number of co-authors, notably appearing via citations!

Advances in Kernel Methods workshop
Gaussian Process and Uncertainty Quantification Summer School
The University of Sheffield, September 6th 2018

Outline

1. Introduction
   - p.d. kernels, from analysis to GPs and back
2. On kernels and invariances
   - Contributions from second order to Gaussian
   - Numerical applications and discussion

Outline

Outline

Kernel methods and invariances/degeneracies

Kernels are a crucial ingredient in a number of mathematical and statistical methods for function approximation, data classification and beyond:

- Support Vector Machines,
- Gaussian Process Modelling,
- Regularization in Reproducing Kernel Hilbert Spaces,
- Kernel Principal Component Analysis,
- Embedding of measures in RKHS,
- Etc.

The implementation of any of these methods require a valid kernel k.

We focus on the choice of k in the function approximation framework and in particular on invariance/degeneracy properties that can be driven by k.

What are (complex- and real-valued) p.d. kernels?

Let D be a set and $k : D \times D \longrightarrow \mathbb{C}$.

k is called a *positive definite* kernel when

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i \overline{a_j} k(\mathbf{x}_i, \mathbf{x}_j) \in [0, +\infty)$$

for all $n \geq 1$, $a_1, \ldots, a_n \in \mathbb{C}$, and $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D$.

What are (complex- and real-valued) p.d. kernels?

Let D be a set and $k : D \times D \longrightarrow \mathbb{C}$.

k is called a *positive definite* kernel when

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i \overline{a_j} k(\mathbf{x}_i, \mathbf{x}_j) \in [0, +\infty)$$

for all $n \geq 1$, $a_1, \ldots, a_n \in \mathbb{C}$, and $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D$.

Follow directly from this definition (More here):

- $k(\mathbf{x}, \mathbf{x}) \in [0, +\infty)$ for all $\mathbf{x} \in D$
- $k(\mathbf{x}', \mathbf{x}) = \overline{k(\mathbf{x}, \mathbf{x}')}$ for all $\mathbf{x}, \mathbf{x}' \in D$ (k is hermitian)
- Non-negative combinations and limits of p.d. kernels are p.d.

What are (complex- and real-valued) p.d. kernels?

Let D be a set and $k : D \times D \longrightarrow \mathbb{C}$.

k is called a *positive definite* kernel when

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i \overline{a_j} k(\mathbf{x}_i, \mathbf{x}_j) \in [0, +\infty)$$

for all $n \geq 1$, $a_1, \ldots, a_n \in \mathbb{C}$, and $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D$.

Follow directly from this definition (More here):

- $k(\mathbf{x}, \mathbf{x}) \in [0, +\infty)$ for all $\mathbf{x} \in D$
- $k(\mathbf{x}', \mathbf{x}) = \overline{k(\mathbf{x}, \mathbf{x}')}$ for all $\mathbf{x}, \mathbf{x}' \in D$ (k is hermitian)
- Non-negative combinations and limits of p.d. kernels are p.d.

NB: $k : D \times D \longrightarrow \mathbb{R}$ is p.d. when both $\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(\mathbf{x}_i, \mathbf{x}_j) \in [0, +\infty)$ for all $n \geq 1$, $a_1, \ldots, a_n \in \mathbb{R}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D$, and k is symmetric.

Considered kernel methods for function approximation

Here we focus on two classes of kernel methods for the approximation of functions based on observational/evaluation data:

- Gaussian Process (GP) modelling/interpolation/regression
- Interpolation/Regularization in Reproducing Kernel Hilbert Spaces

Typical settings of interest are those of an objective function $f : D \longrightarrow \mathbb{R}$ (e.g. with $D \subset \mathbb{R}^d$, $d \geq 1$) that one wishes to approximate relying on a limited number $n \geq 1$ of evaluations at points $\mathbf{x}_i \in D$ ($1 \leq i \leq n$).

About Gaussian Process modelling

GP modelling basically consists in postulating that f is a realization of a real-valued Gaussian random field $Z = (Z_{\mathbf{x}})_{\mathbf{x} \in D}$ and to do inferences on f by using the conditional distribution of Z given the available evaluation results.

About Gaussian Process modelling

GP modelling basically consists in postulating that f is a realization of a real-valued Gaussian random field $Z = (Z_{\mathbf{x}})_{\mathbf{x} \in D}$ and to do inferences on f by using the conditional distribution of Z given the available evaluation results.

As we know, in the Gaussian case the mean and covariance functions (say m and k, here) characterize Z's distribution, so choosing them is crucial.

Reminder: GP/Kriging equations

The GP/Kriging prediction amounts to calculating the conditional expectation and covariance of $Z_{\mathbf{x}}$ knowing $Z_{\mathbf{X}_n} = \mathbf{z}_n$, with $\boldsymbol{z}_n = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))'$:

$$\begin{cases} m_n(\mathbf{x}) = \mathbb{E}[Z_{\mathbf{x}} | Z_{\mathbf{X}_n} = \mathbf{z}_n] = m(\mathbf{x}) + k(\mathbf{x}, \mathbf{X}_n) k(\mathbf{X}_n, \mathbf{X}_n)^{-1} (\mathbf{z}_n - m(\boldsymbol{X}_n)) \\ \\ k_n(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[Z_{\mathbf{x}}, Z_{\mathbf{x}'} | Z_{\mathbf{X}_n} = \mathbf{z}_n] = k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, \mathbf{X}_n) k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x}), \end{cases}$$

Reminder: GP/Kriging equations

The GP/Kriging prediction amounts to calculating the conditional expectation and covariance of $Z_{\mathbf{x}}$ knowing $Z_{\mathbf{X}_n} = \mathbf{z}_n$, with $\boldsymbol{z}_n = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))'$:

$$\begin{cases} m_n(\mathbf{x}) = \mathbb{E}[Z_{\mathbf{x}} | Z_{\mathbf{X}_n} = \mathbf{z}_n] = m(\mathbf{x}) + k(\mathbf{x}, \mathbf{X}_n) k(\mathbf{X}_n, \mathbf{X}_n)^{-1} (\mathbf{z}_n - m(\boldsymbol{X}_n)) \\ \\ k_n(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[Z_{\mathbf{x}}, Z_{\mathbf{x}'} | Z_{\mathbf{X}_n} = \mathbf{z}_n] = k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, \mathbf{X}_n) k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x}), \end{cases}$$

$$\text{where } k(\mathbf{X}_n, \mathbf{X}_n), = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \ldots & k(\mathbf{x}_1, \mathbf{x}_n) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \ldots & k(\mathbf{x}_2, \mathbf{x}_n) \\ \ldots & \ldots & \ldots. & \ldots. \\ k(\mathbf{x}_n, \mathbf{x}_1) & \ldots & \ldots. & k(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix} \text{ and } k(\mathbf{X}_n, \mathbf{x}) = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}) \\ k(\mathbf{x}_2, \mathbf{x}) \\ \ldots \\ k(\mathbf{x}_n, \mathbf{x}) \end{pmatrix}.$$

Reminder: GP/Kriging equations

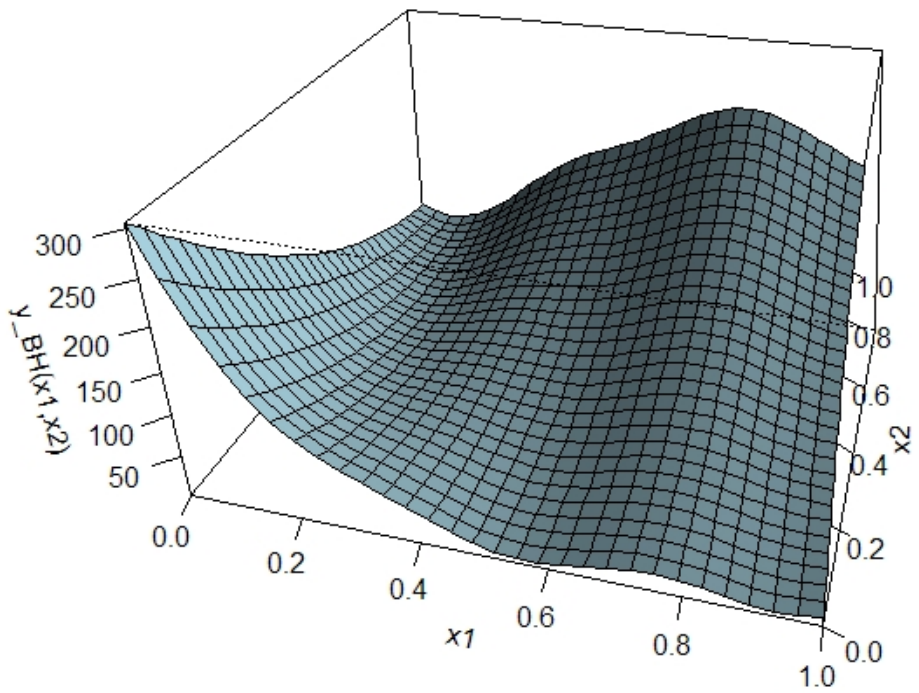
The GP/Kriging prediction amounts to calculating the conditional expectation and covariance of $Z_{\mathbf{x}}$ knowing $Z_{\mathbf{X}_n} = \mathbf{z}_n$, with $\boldsymbol{z}_n = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))'$:

$$\begin{cases} m_n(\mathbf{x}) = \mathbb{E}[Z_{\mathbf{x}} | Z_{\mathbf{X}_n} = \mathbf{z}_n] = m(\mathbf{x}) + k(\mathbf{x}, \mathbf{X}_n) k(\mathbf{X}_n, \mathbf{X}_n)^{-1} (\mathbf{z}_n - m(\boldsymbol{X}_n)) \\ \\ k_n(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[Z_{\mathbf{x}}, Z_{\mathbf{x}'} | Z_{\mathbf{X}_n} = \mathbf{z}_n] = k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, \mathbf{X}_n) k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x}), \end{cases}$$

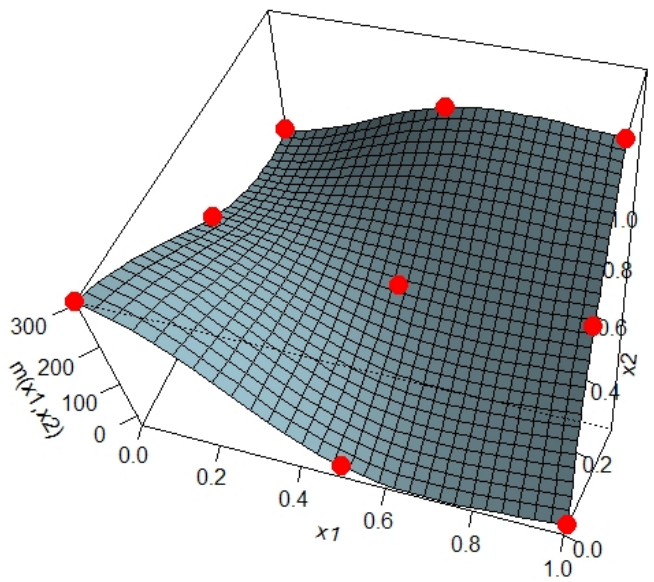
$$\text{where } k(\mathbf{X}_n, \mathbf{X}_n), = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \ldots & k(\mathbf{x}_1, \mathbf{x}_n) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \ldots & k(\mathbf{x}_2, \mathbf{x}_n) \\ \ldots & \ldots & \ldots. & \ldots. \\ k(\mathbf{x}_n, \mathbf{x}_1) & \ldots & \ldots. & k(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix} \text{ and } k(\mathbf{X}_n, \mathbf{x}) = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}) \\ k(\mathbf{x}_2, \mathbf{x}) \\ \ldots \\ k(\mathbf{x}_n, \mathbf{x}) \end{pmatrix}.$$

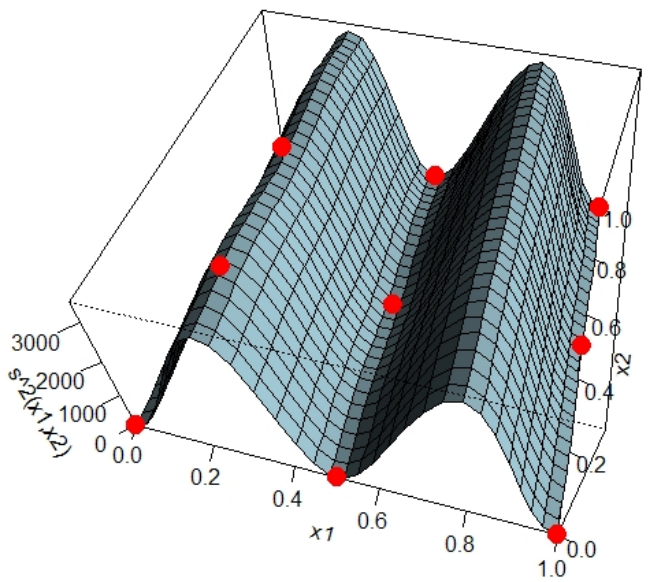
For given m and k (possible generalizations to m known up to linear combination coefficients, cf. Universal Kriging with improper uniform prior), Z knowing $Z_{\mathbf{X}_n} = \mathbf{z}_n$ is a GP with mean m_n and covariance k_n.

A classical test function: Branin-Hoo (Eqs)



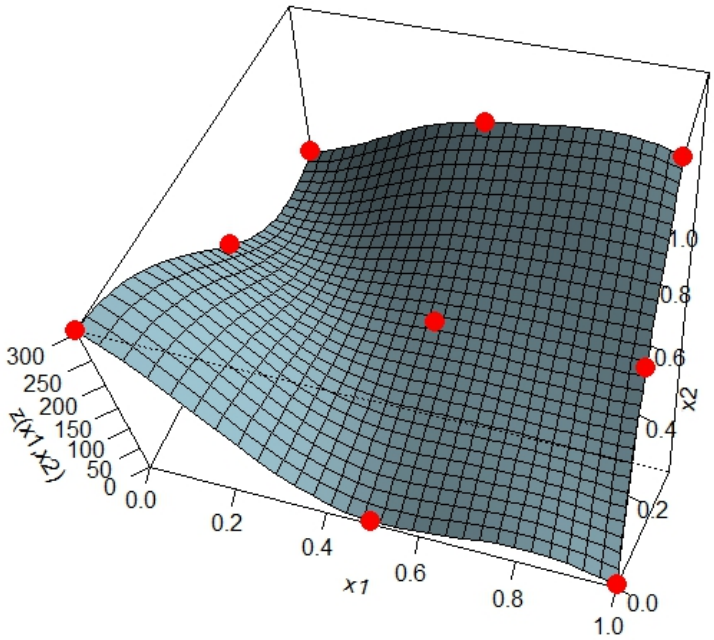
GP Interpolation (Kriging) of the Branin-Hoo function





The covariance is here a **stationary** anisotropic Matérn kernel ($\nu = 5/2$) with scale σ and range parameters (θ_1, θ_2) estimated by Maximum Likelihood.

Conditional simulations of the Branin-Hoo function



A detour through deterministic function approximation

Approximating f based on evaluations at n points is ill-posed without further assumptions on f. Also in deterministic settings, p.d. kernels play a key role.

Kimeldorf, G. and Wahba, G. (1971)

Some results on Tchebycheffian spline functions

Journal of mathematical analysis and applications 33 (1), 82-95

H. Wendland (2005)

Scattered Data Approximation

Cambridge University Press

Fasshauer, G. E. (2011)

Positive definite kernels: past, present and future

Dolomites Research Notes on Approximation, 4:21-63

Scheuerer, M. and Schaback, R. and Schlather, M. (2013)

Interpolation of spatial data - a stochastic or a deterministic problem?

European Journal of Applied Mathematics, 24, 4, 601-629

Optimal approximation in RKHSs

Theorem (Generalization of Kimeldorf and Wahba's 1971's "representer theorem" by Schölkopf, Herbrich and Smola): Given evaluation results

$$(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_n, z_n) \in D \times \mathbb{R},$$

an arbitrary cost function $c : (D \times \mathbb{R}^2)^n \longrightarrow \mathbb{R} \cup \{\infty\}$, and a strictly increasing function p on $[0, \infty)$, any $m_n \in \mathcal{H}_k$ (RKHS with kernel k) minimizing

$$g \in \mathcal{H}_k \longrightarrow c\left((\mathbf{x}_1, z_1, g(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, z_n, g(\mathbf{x}_n))\right) + p(||g||_{\mathcal{H}_k})$$

admits a representation of the form

$$m_n(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, \mathbf{x}_i),$$

with $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ (Notes: noiseless or noisy z_is; real-valued k here.).

Optimal approximation in RKHSs

Theorem (Generalization of Kimeldorf and Wahba's 1971's "representer theorem" by Schölkopf, Herbrich and Smola): Given evaluation results

$$(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_n, z_n) \in D \times \mathbb{R},$$

an arbitrary cost function $c : (D \times \mathbb{R}^2)^n \longrightarrow \mathbb{R} \cup \{\infty\}$, and a strictly increasing function p on $[0, \infty)$, any $m_n \in \mathcal{H}_k$ (RKHS with kernel k) minimizing

$$g \in \mathcal{H}_k \longrightarrow c\left((\mathbf{x}_1, z_1, g(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, z_n, g(\mathbf{x}_n))\right) + p(||g||_{\mathcal{H}_k})$$

admits a representation of the form

$$m_n(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, \mathbf{x}_i),$$

with $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ (Notes: noiseless or noisy z_is; real-valued k here.).

B. Schölkopf, R. Herbrich, A.J. Smola (2001)

A Generalized Representer Theorem

Computational Learning Theory. Lecture Notes in Computer Science 2111:416-426

Outline

In RKHS regularization and GP models with known (e.g., constant) mean, prior assumptions on f are implicitly accounted for through the choice of k.

In RKHS regularization and GP models with known (e.g., constant) mean, prior assumptions on f are implicitly accounted for through the choice of k.

Classical notions of invariance for k

- 2nd order stationarity (k *invariant wrt simult. translations of* $\mathbf{x}$ *and* $\mathbf{x}'$)
- Isotropy (k *invariant wrt simultaneous rigid motions of* $\mathbf{x}$ *and* $\mathbf{x}'$).

In RKHS regularization and GP models with known (e.g., constant) mean, prior assumptions on f are implicitly accounted for through the choice of k.

Classical notions of invariance for k

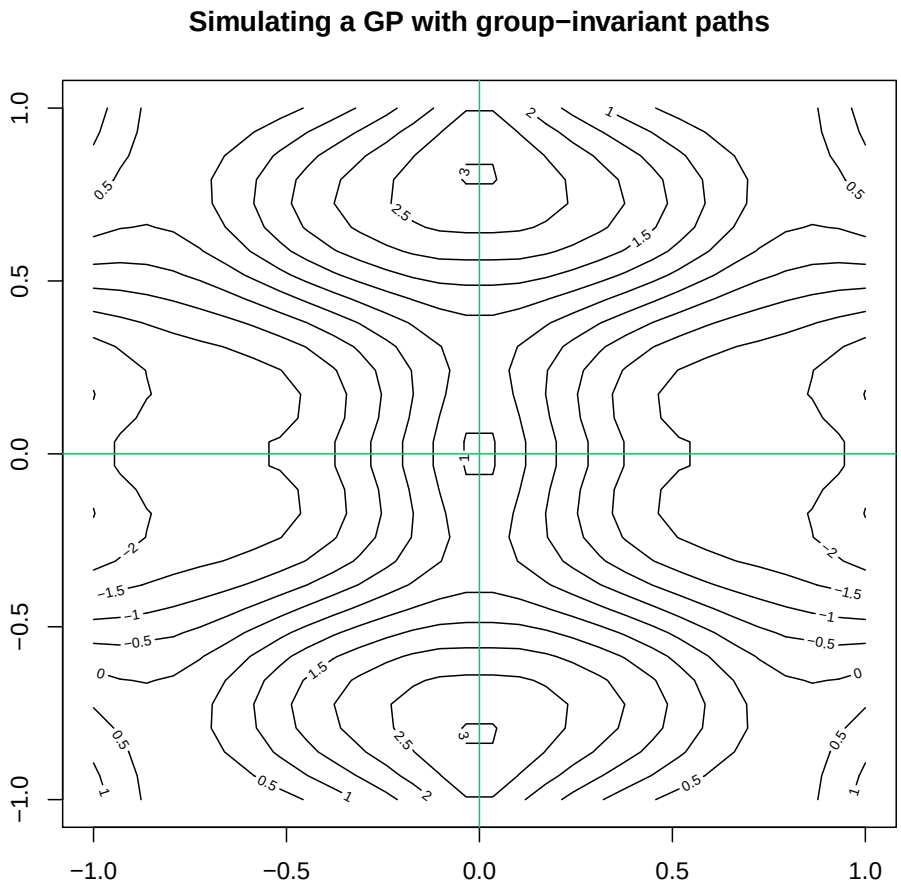
- 2nd order stationarity (k *invariant wrt simult. translations of* $\mathbf{x}$ *and* $\mathbf{x}'$)
- Isotropy (k *invariant wrt simultaneous rigid motions of* $\mathbf{x}$ *and* $\mathbf{x}'$).

We rather investigate some functional properties driven by k, with a main focus on the stochastic case (+ some links to RKHSs).
This talk follows to a large extent the paper below and references therein:

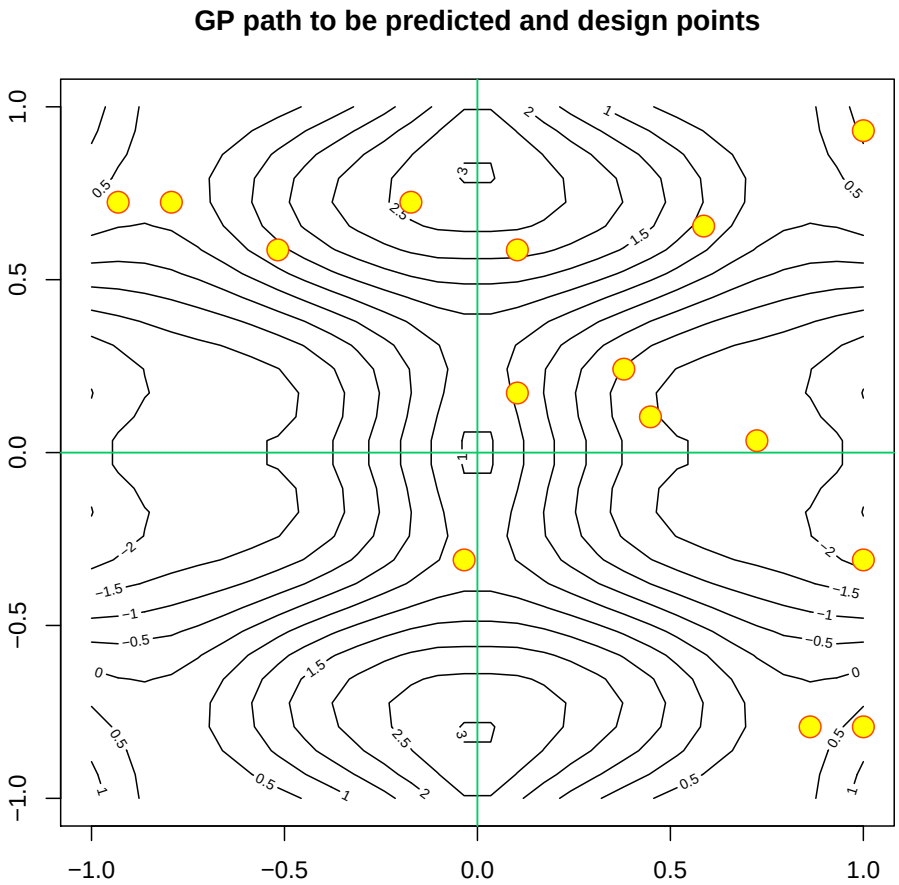
D. G., O. Roustant and N. Durrande (2016)
On degeneracy and invariances of random fields paths with applications in Gaussian Process modelling
Journal of Statistical Planning and Inference, 170:117-128.

Outline

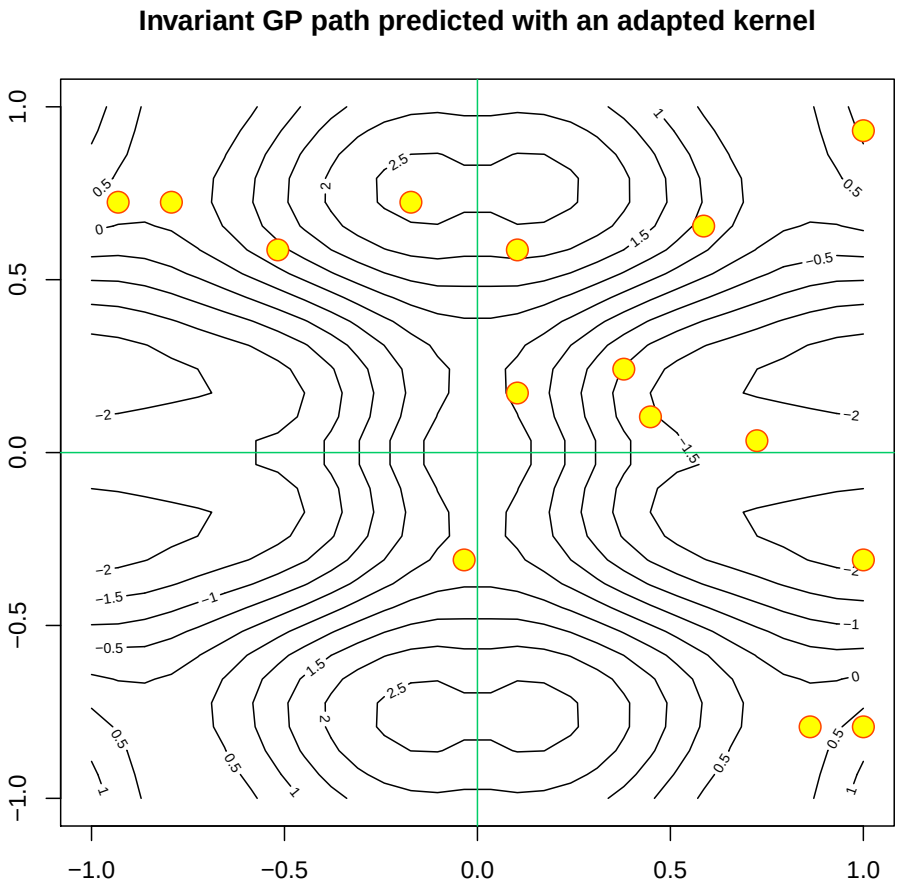
Simulating a GP with group-invariant paths



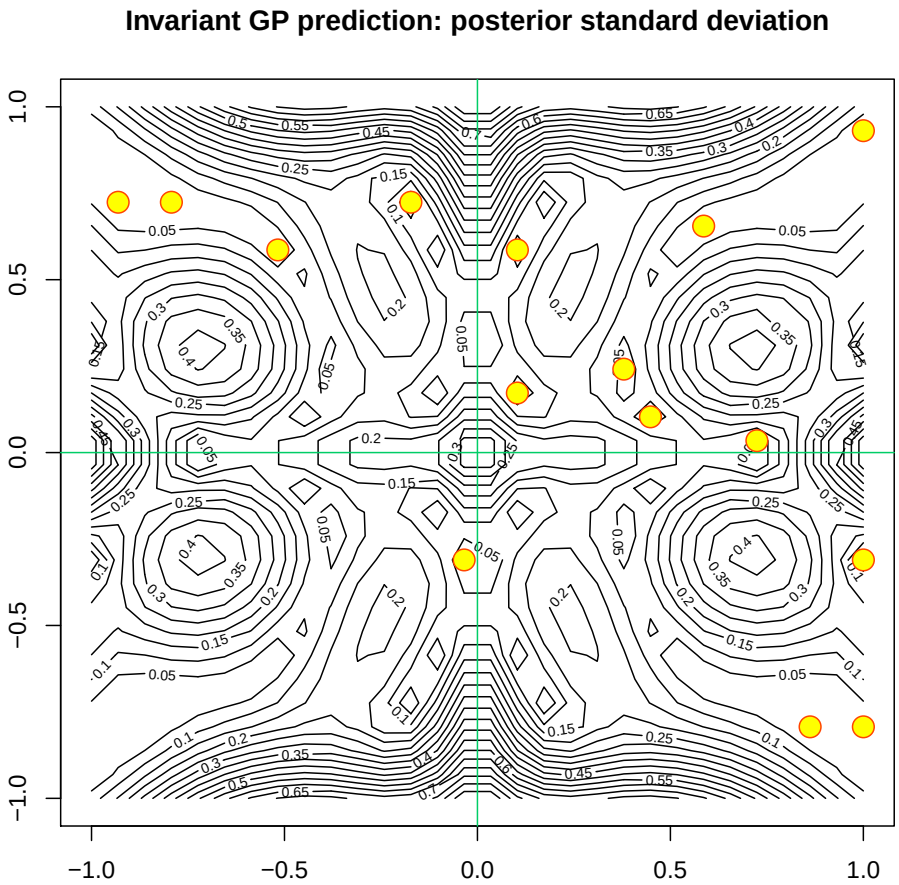
Towards invariant prediction: set-up



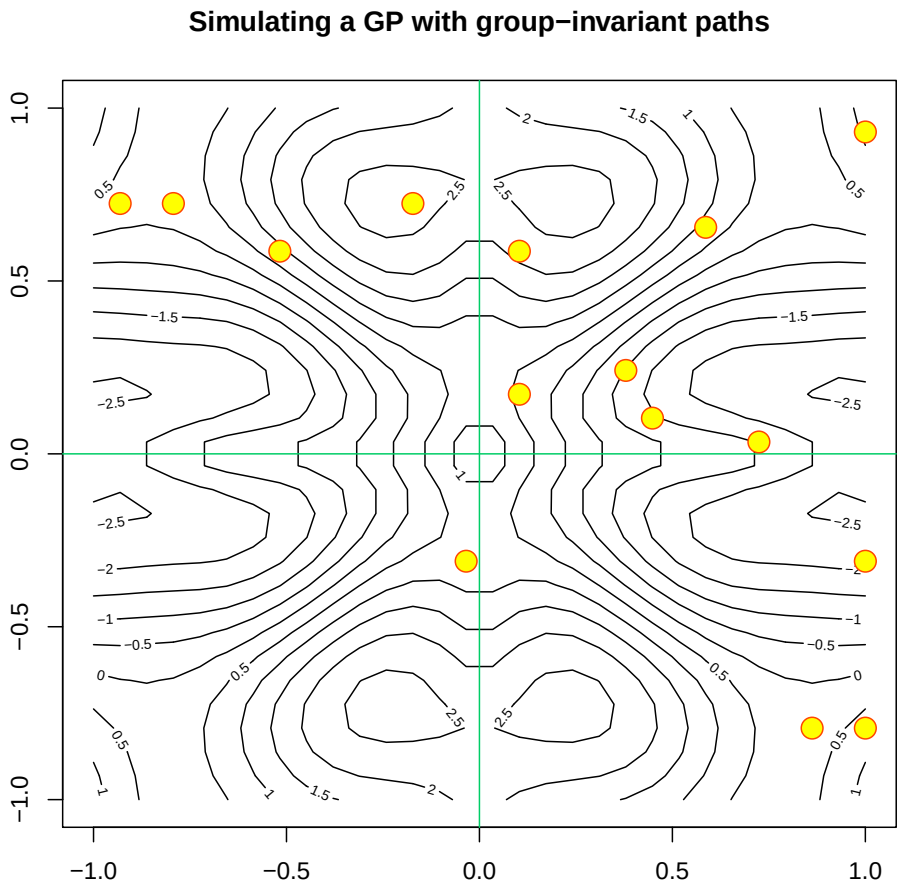
Predicting with an (argumentwise) invariant kernel



Predicting with an (argumentwise) invariant kernel



Invariant conditional simulations



Some refs on group-invariance in kernel methods


B. Haasdonk, H.Burkhardt (2007).
Invariant kernels for pattern analysis and machine learning
Machine Learning 68, 35-61


D. G., X. Bay, O. Roustant and L. Carraro (2012)
Argumentwise invariant kernels for the approximation of invariant functions
Annales de la Faculté des Sciences de Toulouse, 21(3):501-527


K. Hansen et al. (2013)
Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies
Journal of Chemical Theory and Computation 9, 3404-3419


Y. Mroueh, S. Voinea, T. Poggio (2015)
Learning with Group Invariant Features: A Kernel Perspective
Advances in Neural Information Processing Systems, 1558-1566

Proposition (DG et al. 2016)

Let Z be a measurable random field with paths (a.s.) in some function space $\mathcal{F}$ and $T : \mathcal{F} \longrightarrow \mathcal{F}$ be a linear operator such that for all $\mathbf{x} \in D$ there exists a signed measure $\nu_{\mathbf{x}} : \mathcal{D} \longrightarrow \mathbb{R}$ satisfying

$$T(g)(\mathbf{x}) = \int g(\mathbf{u}) \mathrm{d}\nu_{\mathbf{x}}(\mathbf{u}).$$

Assume further that

$$\sup_{\mathbf{x} \in D} \int_D \sqrt{k(\mathbf{u}, \mathbf{u}) + m(\mathbf{u})^2} \mathrm{d}|\nu_{\mathbf{x}}|(\mathbf{u}) < +\infty.$$

Then the following are equivalent:

a) $\forall \mathbf{x} \in D \; \mathbb{P}(T(Z)_{\mathbf{x}} = 0) = 1$ *("$T(Z) = \mathbf{0}$ up to a modification")*

b) $\forall \mathbf{x} \in D \; T(m)(\mathbf{x}) = 0$ *and* $(T \otimes T(k))(\mathbf{x}, \mathbf{x}) = 0$.

Assuming further that $T(Z)$ is separable, ***a)*** *and* ***b)*** *are also equivalent to*

c) $\mathbb{P}(T(Z) = \mathbf{0}) = \mathbb{P}(\forall \mathbf{x} \in D \; T(Z)_{\mathbf{x}} = 0) = 1$ *("$T(Z) = \mathbf{0}$ a.s.") .*

Another invariance: random fields with additive paths

Let $D = \prod_i^d D_i$ where $D_i \subset \mathbb{R}$. $f \in \mathbb{R}^D$ is called additive when there exists $f_i \in \mathbb{R}^{D_i}$ $(1 \leq i \leq d)$ such that $f(\mathbf{x}) = \sum_{i=1}^d f_i(x_i)$ $(\mathbf{x} = (x_1, \ldots, x_d) \in D)$.

Another invariance: random fields with additive paths

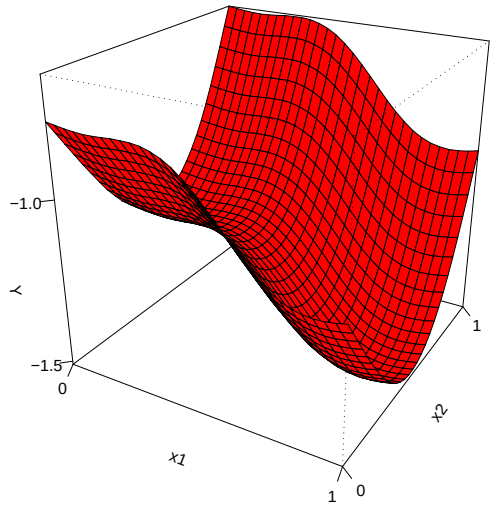
Let $D = \prod_i^d D_i$ where $D_i \subset \mathbb{R}$. $f \in \mathbb{R}^D$ is called additive when there exists $f_i \in \mathbb{R}^{D_i}$ $(1 \leq i \leq d)$ such that $f(\mathbf{x}) = \sum_{i=1}^d f_i(x_i)$ $(\mathbf{x} = (x_1, \ldots, x_d) \in D)$.

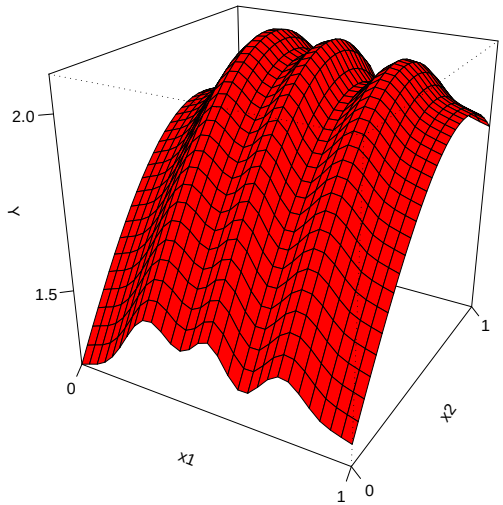
GP models possessing additive paths (with $k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^d k_i(x_i, x_i')$) have been considered in Nicolas Durrande's Ph.D. thesis (2011):

Another invariance: random fields with additive paths

Let $D = \prod_i^d D_i$ where $D_i \subset \mathbb{R}$. $f \in \mathbb{R}^D$ is called additive when there exists $f_i \in \mathbb{R}^{D_i}$ $(1 \leq i \leq d)$ such that $f(\mathbf{x}) = \sum_{i=1}^d f_i(x_i)$ $(\mathbf{x} = (x_1, \ldots, x_d) \in D)$.

GP models possessing additive paths (with $k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^d k_i(x_i, x_i')$) have been considered in Nicolas Durrande's Ph.D. thesis (2011):





A few selected references related to additive kernels


N. Durrande (2011)
Étude de classes de noyaux adaptés à la simplification et à l'interprétation des modèles d'approximation. Une approche fonctionnelle et probabiliste
PhD thesis, Ecole des Mines de Saint-Etienne


D. Duvenaud, H. Nickisch, C. Rasmussen (2011)
Additive Gaussian Processes
Neural Information Processing Systems


N. Durrande, D. G. and O. Roustant (2012)
Additive Covariance kernels for high-dimensional Gaussian Process modeling
Annales de la Faculté des Sciences de Toulouse, 21(3):481-499


D. G., N. Durrande and O. Roustant (2013)
Kernels and designs for modelling invariant functions: From group invariance to additivity.
In mODa 10 - Advances in Model-Oriented Design and Analysis. Contributions to Statistics

A link with RKHSs in the Gaussian case

In Gaussian case, the Loève isometry Ψ between $\mathcal{L}(Z)$ (The Hilbert space generated by Z) and the RKHS $\mathcal{H}_k$ leads to the following.

A link with RKHSs in the Gaussian case

In Gaussian case, the Loève isometry Ψ between $\mathcal{L}(Z)$ (The Hilbert space generated by Z) and the RKHS $\mathcal{H}_k$ leads to the following.

Proposition

Let $T : \mathcal{F} \to \mathbb{R}^D$ be a linear operator such that $T(m) \equiv 0$ and $T(Z)_{\mathbf{x}} \in \mathcal{L}(Z)$ for any $\mathbf{x} \in D$. Then, there exists a unique linear $\mathcal{T} : \mathcal{H}_k \to \mathbb{R}^D$ satisfying

$$\mathrm{cov}(T(Z)_{\mathbf{x}}, Z_{\mathbf{x}'}) = \mathcal{T}(k(\cdot, \mathbf{x}'))(\mathbf{x}) \quad (\mathbf{x}, \mathbf{x}' \in D)$$

and such that $\mathcal{T}(h_n)(\mathbf{x}) \longrightarrow \mathcal{T}(h)(\mathbf{x})$ for any $\mathbf{x} \in D$ and $h_n \xrightarrow{\mathcal{H}} h$.

In addition, we have equivalence between the following:

(i) $\forall \mathbf{x} \in D \; T(Z)_{\mathbf{x}} = 0$ *(almost surely)*

(iii) $\forall \mathbf{x}' \in D \; \mathcal{T}(k(\cdot, \mathbf{x}')) = \mathbf{0}$

(iii) $\mathcal{T}(\mathcal{H}_k) = \{0\}$

Examples

a) Let ν be a measure on D s.t. $\int_D \sqrt{k(\mathbf{u},\mathbf{u})}d\nu(\mathbf{u}) < +\infty$. Then a centred Z (Gaussian or not) has centred paths iff $\int_D k(\mathbf{x},\mathbf{u})\mathrm{d}\nu(\mathbf{u}) = 0, \forall \mathbf{x} \in D$.

For instance, given any p.d. kernel k, k_0 defined by

$$k_0(\mathbf{x},\mathbf{y}) = k(\mathbf{x},\mathbf{y}) - \int k(\mathbf{x},\mathbf{u})\mathrm{d}\nu(\mathbf{u}) - \int k(\mathbf{y},\mathbf{u})\mathrm{d}\nu(\mathbf{u}) + \int k(\mathbf{u},\mathbf{v})\mathrm{d}\nu(\mathbf{u})\mathrm{d}\nu(\mathbf{v})$$

satisfies the above condition.

Examples

a) Let ν be a measure on D s.t. $\int_D \sqrt{k(\mathbf{u},\mathbf{u})} d\nu(\mathbf{u}) < +\infty$. Then a centred Z (Gaussian or not) has centred paths iff $\int_D k(\mathbf{x},\mathbf{u}) \mathrm{d}\nu(\mathbf{u}) = 0, \forall \mathbf{x} \in D$.

For instance, given any p.d. kernel k, k_0 defined by

$$k_0(\mathbf{x},\mathbf{y}) = k(\mathbf{x},\mathbf{y}) - \int k(\mathbf{x},\mathbf{u}) \mathrm{d}\nu(\mathbf{u}) - \int k(\mathbf{y},\mathbf{u}) \mathrm{d}\nu(\mathbf{u}) + \int k(\mathbf{u},\mathbf{v}) \mathrm{d}\nu(\mathbf{u}) \mathrm{d}\nu(\mathbf{v})$$

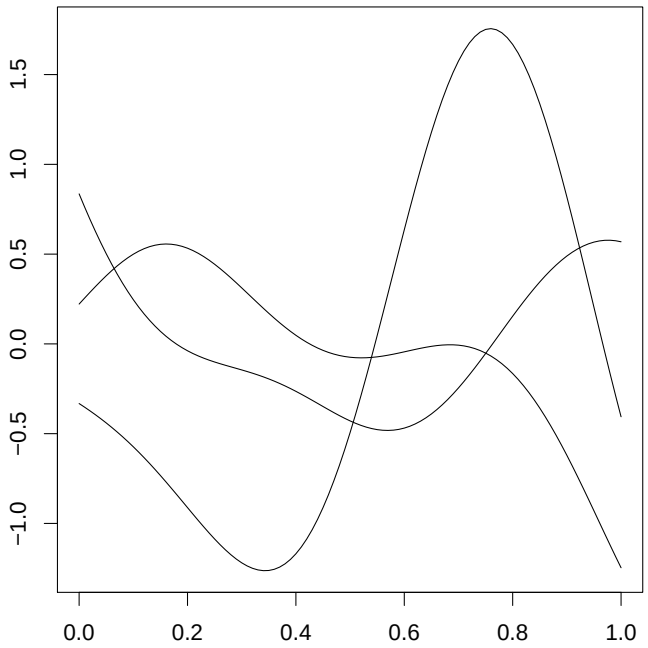
satisfies the above condition.

b) Solutions to the *Laplace equation* are called harmonic functions. Let us call harmonic any p.d. kernel solving the Laplace equation argumentwise: $(\Delta k(\cdot, \mathbf{x}')) = 0$ $(\mathbf{x}' \in D)$.

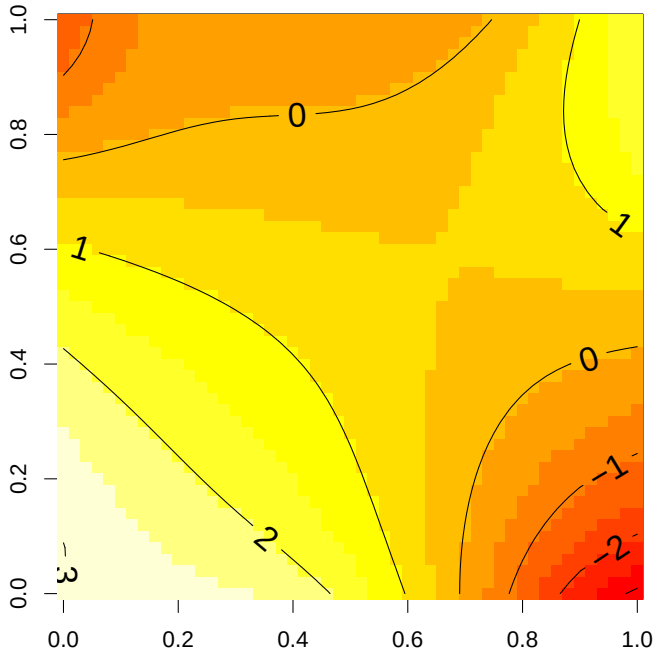
An example of such harmonic kernel over $\mathbb{R}^2 \times \mathbb{R}^2$ can be found in the recent literature (Schaback et al. 2009):

$$k_{harm}(\mathbf{x},\mathbf{y}) = \exp\left(\frac{x_1 y_1 + x_2 y_2}{\theta^2}\right) \cos\left(\frac{x_2 y_1 - x_1 y_2}{\theta^2}\right).$$

Example sample paths invariant under various T's



(a) Zero-mean paths of the centred GP with kernel k_0.



(b) Harmonic path of a GRF with kernel k_{harm}.

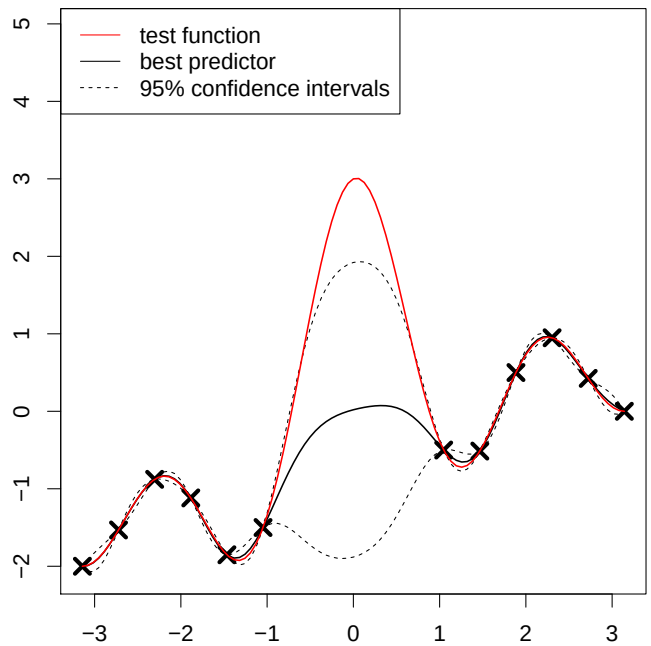
Some “stability of invariances by conditioning” result

Proposition

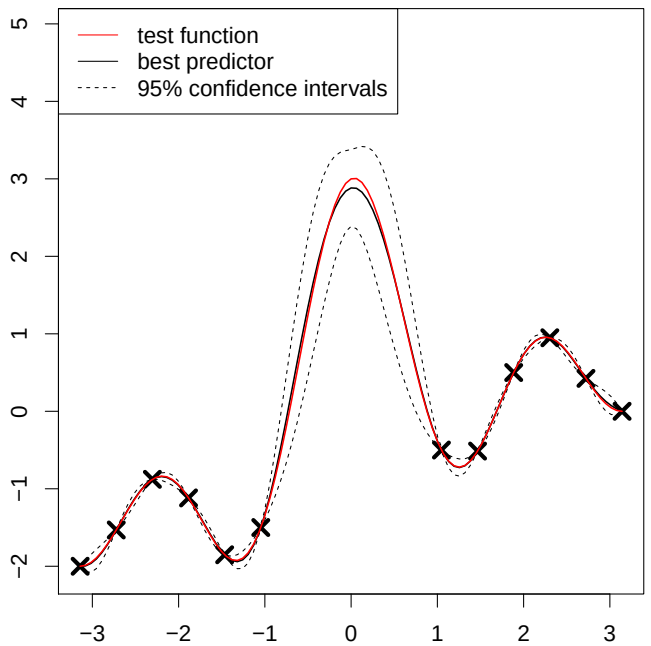
- *Let $\mathcal{F}, \mathcal{G}$ be real separable Banach spaces,*
- *μ be a Gaussian measure on $\mathcal{B}(\mathcal{F})$ with mean zero and covariance operator C_μ*
- *$T : \mathcal{F} \longrightarrow \mathcal{F}$ be a bounded linear operator such that $TC_\mu T^\star = 0_{\mathcal{F}^\star \longrightarrow \mathcal{F}}$*
- *$A : \mathcal{F} \longrightarrow \mathcal{G}$ be another bounded linear operator,*
- *and $A_\sharp \mu$ be the image of μ under A.*

Then there exist a Borel measurable mapping $m : \mathcal{G} \longrightarrow \mathcal{F}$, a Gaussian covariance $R : \mathcal{F}^\star \longrightarrow \mathcal{F}$ with $R \leq C_\mu$ and a disintegration $(q_y)_{y \in \mathcal{G}}$ of μ on $\mathcal{B}(\mathcal{F})$ with respect to A such that for any fixed $y \in \mathcal{G}$, q_y is a Gaussian measure with mean m and covariance operator R satisfying $T(m) = 0_{\mathcal{F}}$ and $TRT^\star = 0_{\mathcal{F}^\star \longrightarrow \mathcal{F}}$.

GP prediction with invariant kernels: example a)



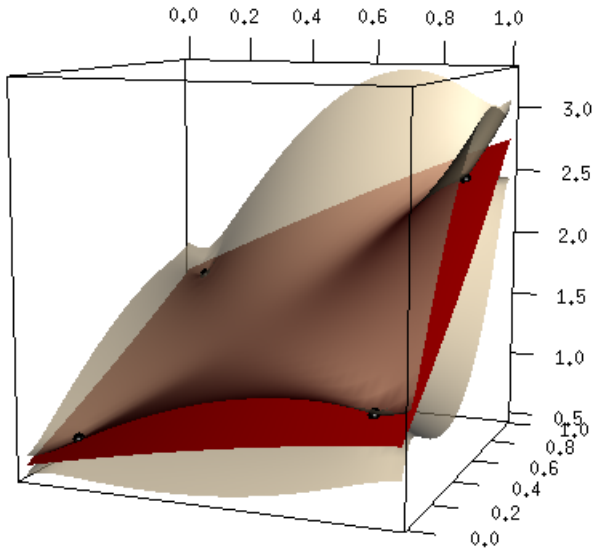
(a) GPR with kernel k



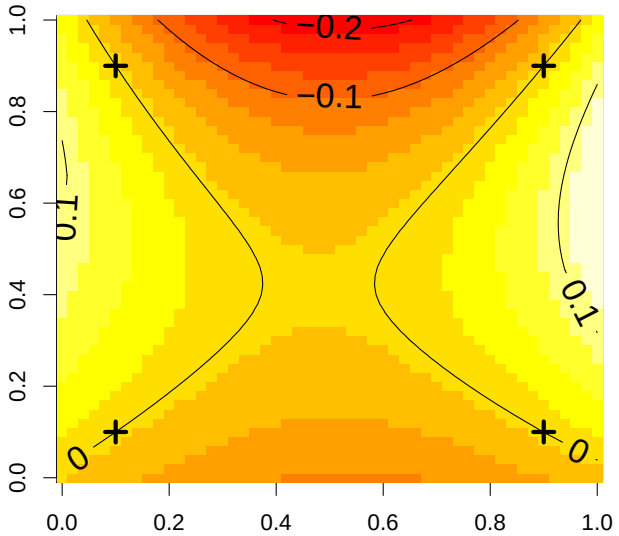
(b) GPR with kernel k_0

Figure: Comparison of two GP models. The left one is based on a Gaussian kernel. The right one incorporates the zero-mean property.

GP models with invariant kernels: example b)



(a) Mean predictor and 95% prediction intervals



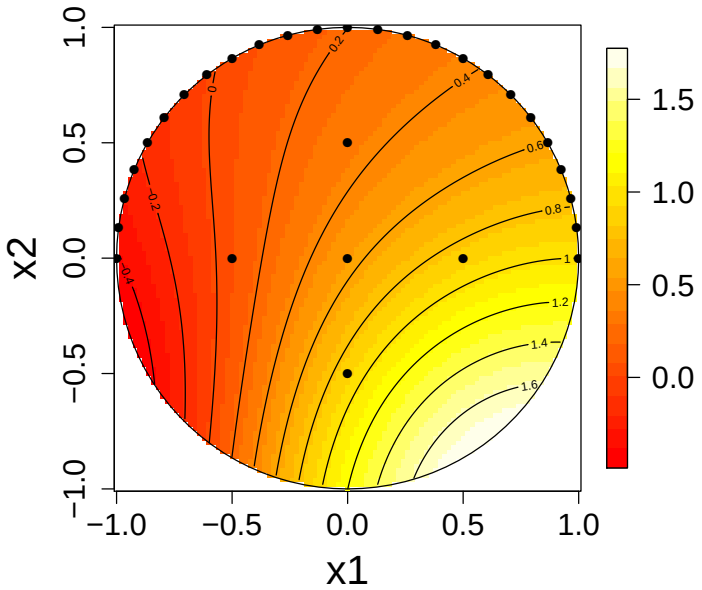
(b) prediction error

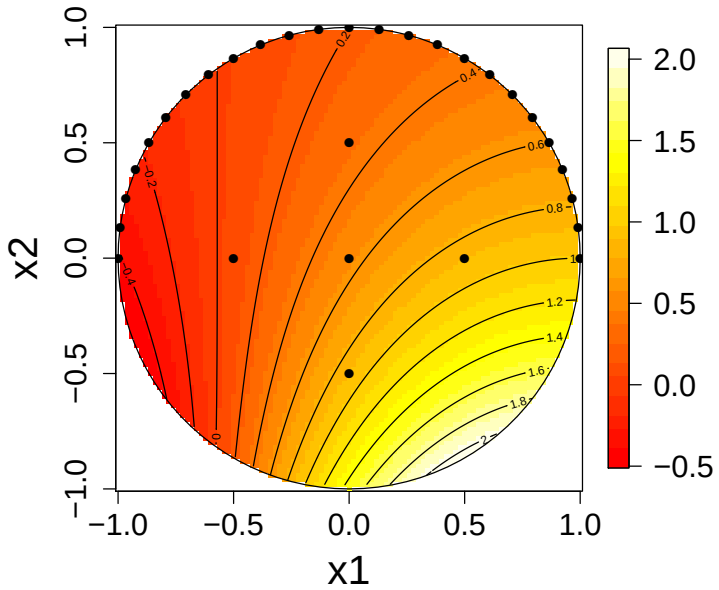
Figure: Example of GP model based on a harmonic kernel.

Outline

Numerical application: maximum of a harmonic *f*

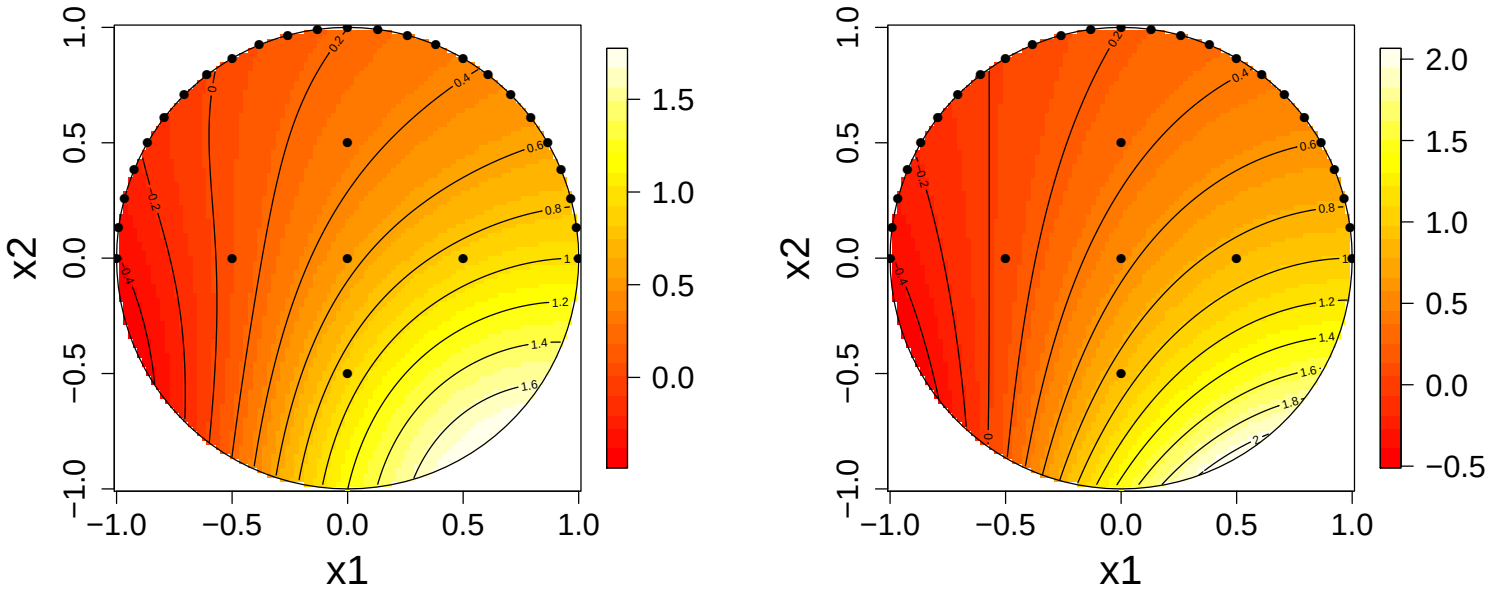
Here we consider approximating a harmonic function (left/right: Gaussian/harmonic kernels) and estimating its maximum by GRF modelling.





Numerical application: maximum of a harmonic *f*

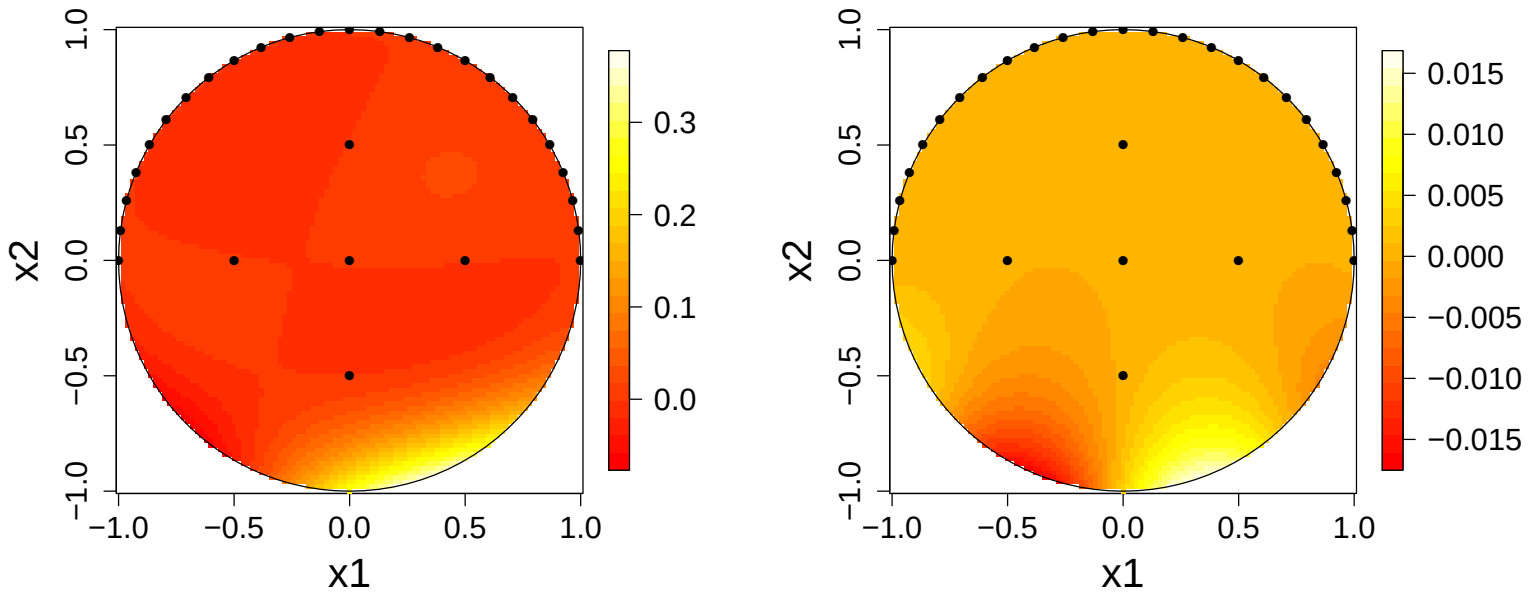
Here we consider approximating a harmonic function (left/right: Gaussian/harmonic kernels) and estimating its maximum by GRF modelling.



Extracted from “On degeneracy and invariances of random fields paths with applications in Gaussian Process modelling” (DG, O.Roustant & N.Durrande, Journal of Statistical Planning and Inference, 170:117-128, 2016)

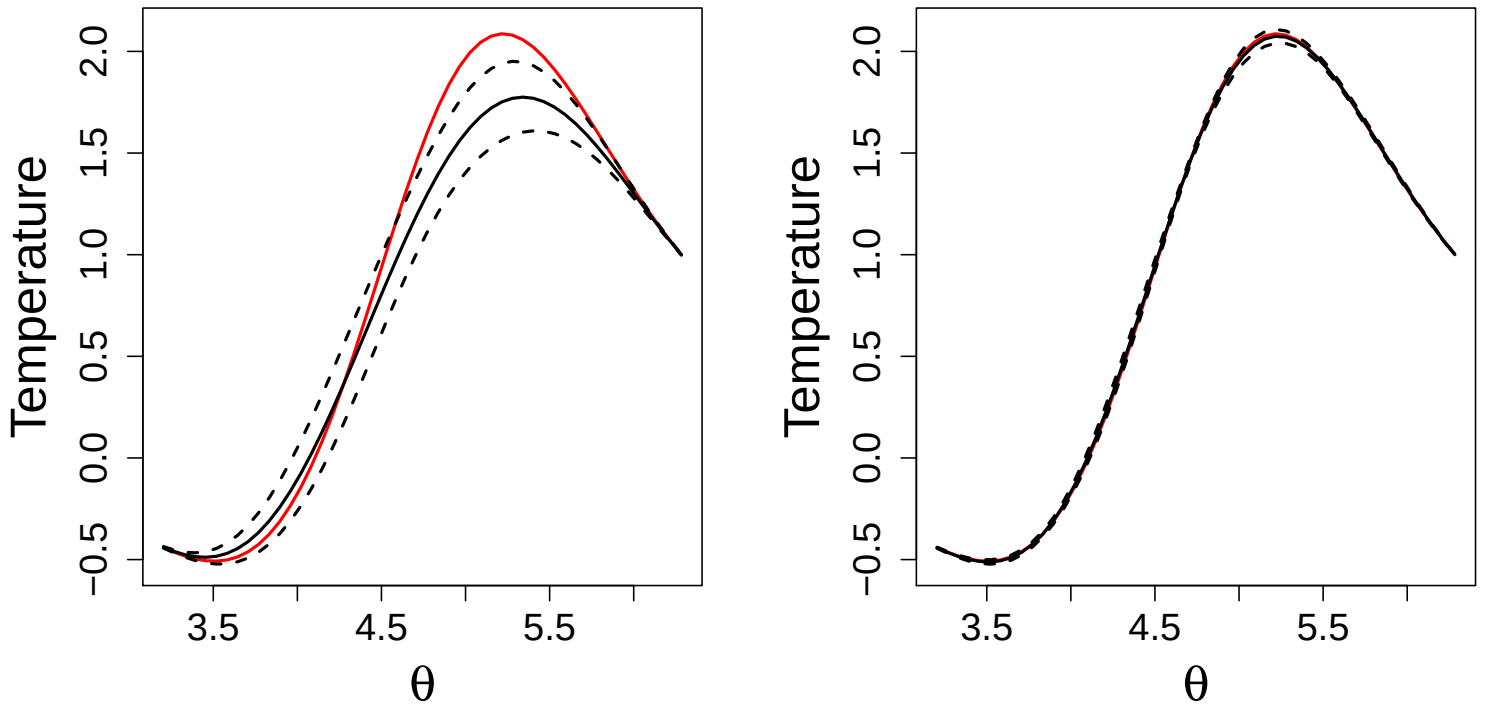
Numerical application: maximum of a harmonic *f*

Prediction errors (left/right: Gaussian/harmonic kernels).



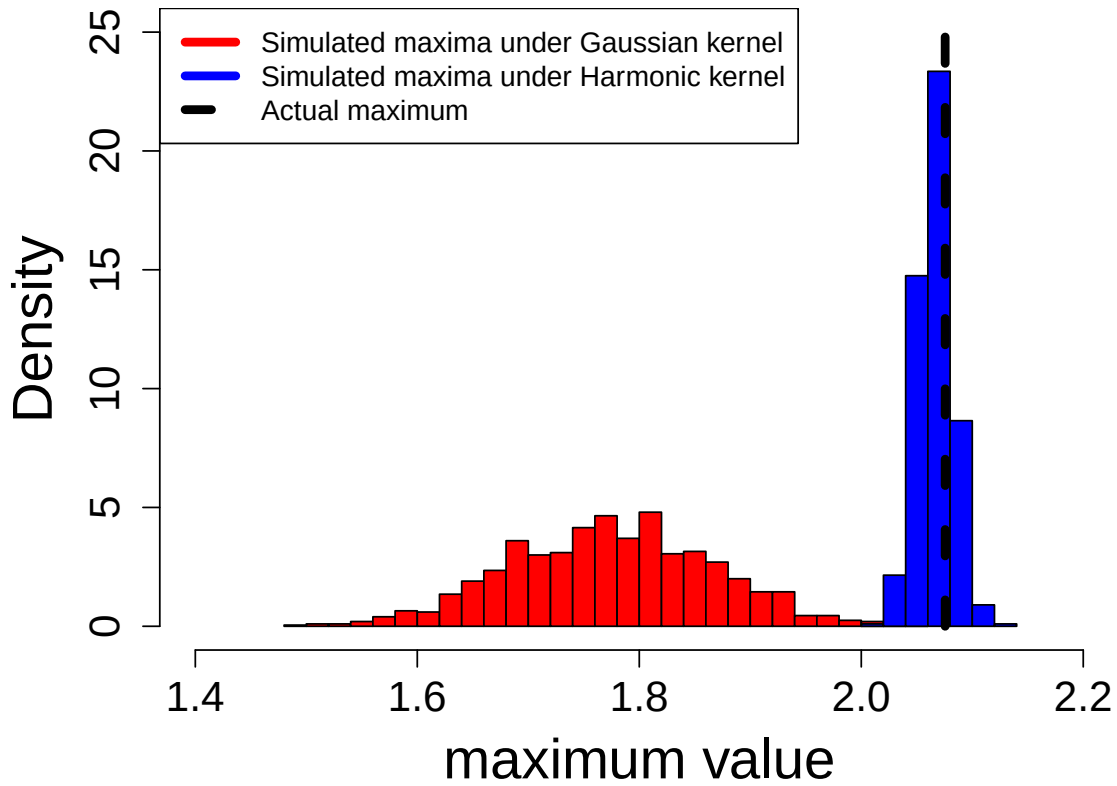
Numerical application: maximum of a harmonic *f*

Prediction errors (left/right: Gaussian/harmonic kernels).

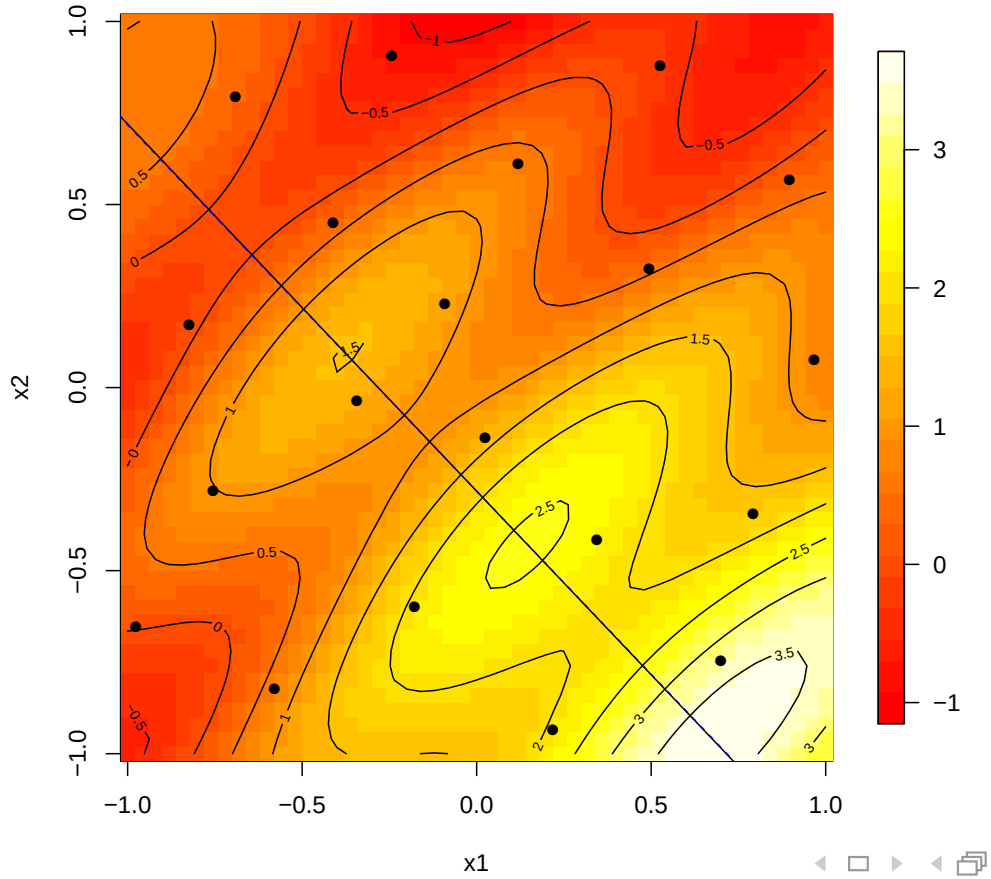


Numerical application: maximum of a harmonic *f*

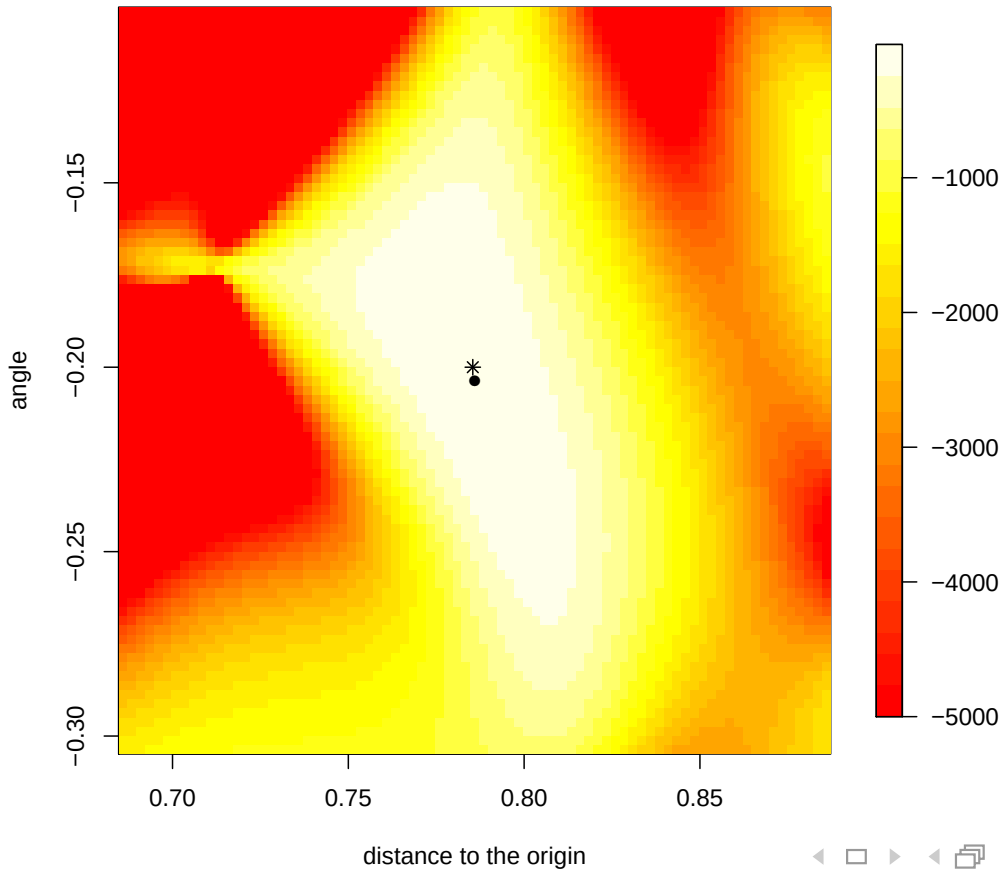
Conditional simulations of the maximum under the two GRF models.



Numerical application: recovering a symmetry axis



Numerical application 2: recovering a symmetry axis



Discussion

Function approximation approaches based on p.d. kernels enable incorporating degeneracies and invariances under linear operators including

- Symmetries and further invariances under group actions
- Additivity and further multivariate sparsity properties towards high-dimensional GRF modelling (See, e.g., MCQMC2014 paper)
- Harmonicity but also, e.g., divergence-free properties for vector fields (See, e.g., Scheuerer and Schlather 2012)

In the Gaussian set up, such properties are inherited by conditional distributions, which is clearly convenient but also comes withs risks.

Perspectives

- Developing further the inference of degeneracy/invariance properties based on data and investigating consistency,

Perspectives

- Developing further the inference of degeneracy/invariance properties based on data and investigating consistency,
- Creating classes of kernels that incorporate some invariant components and non-invariant components,

Perspectives

- Developing further the inference of degeneracy/invariance properties based on data and investigating consistency,
- Creating classes of kernels that incorporate some invariant components and non-invariant components,
- Explore further the potential of invariant kernels based on real-world applications (e.g., from physics, biology, engineering).

Thank you very much for your attention!

Further references


C.J. Stone (1985)
Additive regression and other nonparametric models
The Annals of Statistics 13(2):689-705


M. Scheuerer and M. Schlather (2012)
Covariance Models for Divergence-Free and Curl-Free Random Vector Fields
Stochastic Models 28(3)


D. Duvenaud (2014)
Automatic Model Construction with Gaussian Processes
PhD thesis, University of Cambridge


K. Kandasamy, J. Schneider and B. Poczos (2015)
High Dimensional Bayesian Optimisation and Bandits via Additive Models
International Conference on Machine Learning (ICML) 2015


D. G., O. Roustant, D. Schuhmacher, N. Durrande and N. Lenz (2016)
On ANOVA decompositions of kernels and Gaussian random field paths.
Monte Carlo and Quasi-Monte Carlo Methods

Appendix back

Outline

What do we assume about f in GP modelling?

In Gaussian Process (GP) modelling, probabilistic concepts are used to model the deterministic function f.

What do we assume about f in GP modelling?

In Gaussian Process (GP) modelling, probabilistic concepts are used to model the deterministic function f.

Let us first focus on an arbitrary point $\mathbf{x} \in D$ and think of the unknown response value $f(\mathbf{x})$ as a Gaussian random variable, denoted here $Z_{\mathbf{x}}$.

Of course, how the mean and variance of $Z_{\mathbf{x}}$ are specified is crucial. A simple option is to set them to constant values (e.g. 0 mean and $\sigma^2 > 0$ variance) ...

What do we assume about f in GP modelling?

In Gaussian Process (GP) modelling, probabilistic concepts are used to model the deterministic function f.

Let us first focus on an arbitrary point $\mathbf{x} \in D$ and think of the unknown response value $f(\mathbf{x})$ as a Gaussian random variable, denoted here $Z_{\mathbf{x}}$.

Of course, how the mean and variance of $Z_{\mathbf{x}}$ are specified is crucial. A simple option is to set them to constant values (e.g. 0 mean and $\sigma^2 > 0$ variance) ...

... However, a white noise assumption would not be very constructive! The crux in GP modelling is to assume that the $Z_{\mathbf{x}}$'s for different $\mathbf{x}$'s are correlated.

Reminder: n-dimensional Gaussian distribution

More precisely, we will appeal to the multivariate Gaussian distribution. Let us forget about $\mathbf{x}$ for now and consider a random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$.

Reminder: n-dimensional Gaussian distribution

More precisely, we will appeal to the multivariate Gaussian distribution. Let us forget about $\mathbf{x}$ for now and consider a random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$.

$\mathbf{Z}$ is said to be multivariate Gaussian distributed when $\sum_{i=1}^{n} a_i Z_i$ is Gaussian distributed whatever $n \geq 1$ and $a_1, \ldots, a_n \in \mathbb{R}$.

Reminder: n-dimensional Gaussian distribution

More precisely, we will appeal to the multivariate Gaussian distribution. Let us forget about $\mathbf{x}$ for now and consider a random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$.

$\mathbf{Z}$ is said to be multivariate Gaussian distributed when $\sum_{i=1}^{n} a_i Z_i$ is Gaussian distributed whatever $n \geq 1$ and $a_1, \ldots, a_n \in \mathbb{R}$.

Such $\mathbf{Z}$ is characterized by its mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $K \in \mathbb{R}^{n \times n}$ (with $\mathbb{E}[Z_i]$ and $\mathrm{Cov}[Z_i, Z_j] = \mathbb{E}[(Z_i - \mu_i)(Z_j - \mu_j)]$ entries, respectively).

Reminder: n-dimensional Gaussian distribution

More precisely, we will appeal to the multivariate Gaussian distribution. Let us forget about $\mathbf{x}$ for now and consider a random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$.

$\mathbf{Z}$ is said to be multivariate Gaussian distributed when $\sum_{i=1}^{n} a_i Z_i$ is Gaussian distributed whatever $n \geq 1$ and $a_1, \ldots, a_n \in \mathbb{R}$.

Such $\mathbf{Z}$ is characterized by its mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $K \in \mathbb{R}^{n \times n}$ (with $\mathbb{E}[Z_i]$ and $\mathrm{Cov}[Z_i, Z_j] = \mathbb{E}[(Z_i - \mu_i)(Z_j - \mu_j)]$ entries, respectively).
We use the notation:

$$\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, K).$$

Note that while μ can take any value, K must be symmetric positive semi-definite (i.e. symmetric with non-negative eigenvalues).

Reminder: n-dimensional Gaussian distribution

In case of invertible K, $\boldsymbol{Z}$ possesses the probability density function:

$$p_{\mathcal{N}(\boldsymbol{\mu},K)}(\boldsymbol{z}) = (2\pi)^{-n/2} \det(K)^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})' K^{-1} (\boldsymbol{z}-\boldsymbol{\mu})\right)$$

Reminder: n-dimensional Gaussian distribution

In case of invertible K, $\boldsymbol{Z}$ possesses the probability density function:

$$p_{\mathcal{N}(\boldsymbol{\mu},K)}(\boldsymbol{z}) = (2\pi)^{-n/2}\det(K)^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})'K^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right)$$

Besides that, denoting by $\boldsymbol{Z}_a$ and $\boldsymbol{Z}_b$ two subvectors of $\boldsymbol{Z}$ such that $\boldsymbol{Z} = (\boldsymbol{Z}_a, \boldsymbol{Z}_b)$, by $\boldsymbol{\mu}_a, \boldsymbol{\mu}_b$ the corresponding means, and defining the corresponding blocks of $\boldsymbol{Z}$'s covariance matrix by

$$K = \begin{pmatrix} K_a & K_{ab} \\ K_{ba} & K_b \end{pmatrix},$$

then (assuming invertibility of K_a), the conditional probability distribution of $\boldsymbol{Z}_b$ knowing that $\boldsymbol{Z}_a = \mathbf{z}_a$ is (multivariate) Gaussian with

$$\mathcal{L}(\boldsymbol{Z}^{(b)}|\boldsymbol{Z}_a = \mathbf{z}_a) = \mathcal{N}(\boldsymbol{\mu}_b + K_{ba}K_a^{-1}(\mathbf{z}_a - \boldsymbol{\mu}_a), K_b - K_{ba}K_a^{-1}K_{ab}).$$

Priors on functions?

Let us now come back to our function approximation problem. We are interested in having a prior distribution on functions, not just on a finite-dimensional vector!

Priors on functions?

Let us now come back to our function approximation problem. We are interested in having a prior distribution on functions, not just on a finite-dimensional vector!

Good news from probability theory (Kolmogorov's extension theorem): random processes on D (a.k.a. random fields in case of multivariate D) can be defined through finite-dimensional distributions, i.e. through distributions of the random vectors $(Z_{\mathbf{x}_1}, \ldots, Z_{\mathbf{x}_n})$ for any finite set of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Priors on functions?

Let us now come back to our function approximation problem. We are interested in having a prior distribution on functions, not just on a finite-dimensional vector!

Good news from probability theory (Kolmogorov's extension theorem): random processes on D (a.k.a. random fields in case of multivariate D) can be defined through finite-dimensional distributions, i.e. through distributions of the random vectors $(Z_{\mathbf{x}_1}, \ldots, Z_{\mathbf{x}_n})$ for any finite set of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Gaussian Processes (a.k.a. Gaussian Random Fields)

A GP (GRF) Z with index set D is a collection of random variables $(Z_{\mathbf{x}})_{\mathbf{x} \in D}$ (defined over the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$) such that for any finite set of points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D$, $(Z_{\mathbf{x}_1}, \ldots, Z_{\mathbf{x}_n})$ is multivariate Gaussian

Mean and covariance functions of a GP

Hence a GP is Z defined by specifying the mean and the covariance matrix of any random vector of the form $(Z_{\mathbf{x}_1}, \ldots, Z_{\mathbf{x}_n})$, so that Z is characterized by

$$\mu : \mathbf{x} \in D \longrightarrow \mu(\mathbf{x}) = \mathbb{E}[Z_{\mathbf{x}}] \in \mathbb{R}$$

$$k : (\mathbf{x}, \mathbf{x}') \in D \times D \longrightarrow k(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[Z_{\mathbf{x}}, Z_{\mathbf{x}'}] \in \mathbb{R}$$

Mean and covariance functions of a GP

Hence a GP is Z defined by specifying the mean and the covariance matrix of any random vector of the form $(Z_{\mathbf{x}_1}, \ldots, Z_{\mathbf{x}_n})$, so that Z is characterized by

$$\mu : \mathbf{x} \in D \longrightarrow \mu(\mathbf{x}) = \mathbb{E}[Z_{\mathbf{x}}] \in \mathbb{R}$$
$$k : (\mathbf{x}, \mathbf{x}') \in D \times D \longrightarrow k(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[Z_{\mathbf{x}}, Z_{\mathbf{x}'}] \in \mathbb{R}$$

While μ can be any function, k is constrained since $(k(\mathbf{x}_i, \mathbf{x}_j))_{1 \leq i \leq n, 1 \leq j \leq n}$ must be symmetric positive semi-definite for any set of points.

k satisfying such property are referred to as **p.d. kernels**.

Mean and covariance functions of a GP

Hence a GP is Z defined by specifying the mean and the covariance matrix of any random vector of the form $(Z_{\mathbf{x}_1}, \ldots, Z_{\mathbf{x}_n})$, so that Z is characterized by

$$\mu : \mathbf{x} \in D \longrightarrow \mu(\mathbf{x}) = \mathbb{E}[Z_{\mathbf{x}}] \in \mathbb{R}$$
$$k : (\mathbf{x}, \mathbf{x}') \in D \times D \longrightarrow k(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[Z_{\mathbf{x}}, Z_{\mathbf{x}'}] \in \mathbb{R}$$

While μ can be any function, k is constrained since $(k(\mathbf{x}_i, \mathbf{x}_j))_{1 \leq i \leq n, 1 \leq j \leq n}$ must be symmetric positive semi-definite for any set of points.

k satisfying such property are referred to as **p.d. kernels**.

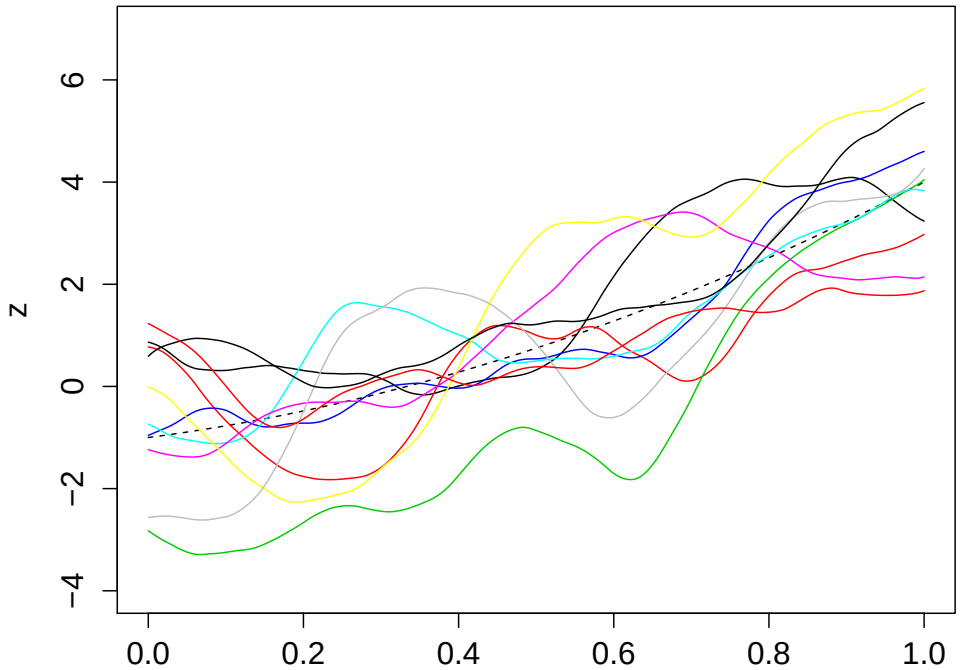
Remark: Assuming $\mu \equiv 0$ for now, k accounts for a number of properties of Z, including *pathwise properties*, i.e. functional properties of the paths

$$\mathbf{x} \in D \longrightarrow Z_{\mathbf{x}}(\omega) \in \mathbb{R},$$

for $\omega \in \Omega$ (paths are also called "realizations", or "trajectories").

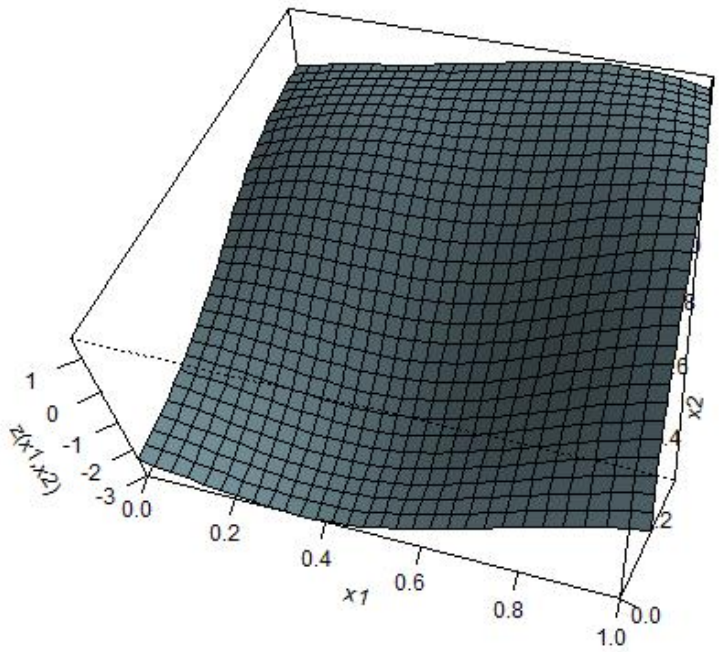
Some GRF R simulations (d=1) with *DiceKriging*

Here $k(t,t') = \sigma^2 \left(1 + |t'-t|/\ell + (t-t')^2/(3\ell^2)\right) \exp\left(-|t'-t|/\ell\right)$
(*Matérn kernel* with regularity parameter $5/2$) where $\ell = 0.4$ and $\sigma = 1.5$.
Furthermore, here trend is a trend $\mu(t) = -1 + 2t + 3t^2$.



Some GRF R simulations (d=2) with *DiceKriging*

Now take a tensorized version of Matérn kernel and a constant trend $\mu = 0$.



Approximating functions using GP models

Let us now consider a deterministic function $f : D \longrightarrow \mathbb{R}$, which response values are measured at n points $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in D^n$.

Putting a GP prior Z on f and updating it with respect to f's values at the $\mathbf{x}_i$ points, we can work out a posterior distribution.

Approximating functions using GP models

Let us now consider a deterministic function $f : D \longrightarrow \mathbb{R}$, which response values are measured at n points $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in D^n$.

Putting a GP prior Z on f and updating it with respect to f's values at the $\mathbf{x}_i$ points, we can work out a posterior distribution.

Indeed, finite-dimensional distributions of this posterior can be obtained by looking at the conditional distribution of $(Z_{\mathbf{x}_{n+1}}, \ldots, Z_{\mathbf{x}_{n+q}})$ knowing $(Z_{\mathbf{x}_1}, \ldots, Z_{\mathbf{x}_n})$ for arbitrary points $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+q} \in D$.

By Gaussianity, it turns out that such conditional distributions are Gaussian and so the posterior $Z|$ measurements is a GRF.

Approximating functions using GP models

Let us now consider a deterministic function $f : D \longrightarrow \mathbb{R}$, which response values are measured at n points $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in D^n$.

Putting a GP prior Z on f and updating it with respect to f's values at the $\mathbf{x}_i$ points, we can work out a posterior distribution.

Indeed, finite-dimensional distributions of this posterior can be obtained by looking at the conditional distribution of $(Z_{\mathbf{x}_{n+1}}, \ldots, Z_{\mathbf{x}_{n+q}})$ knowing $(Z_{\mathbf{x}_1}, \ldots, Z_{\mathbf{x}_n})$ for arbitrary points $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+q} \in D$.

By Gaussianity, it turns out that such conditional distributions are Gaussian and so the posterior $Z|$ measurements is a GRF.

NB: the same applied in noisy cases when considering $(Z_{\mathbf{x}_1} + \varepsilon_1, \ldots, Z_{\mathbf{x}_n} + \varepsilon_n)$ with Gaussian ε_i's independent of Z).

About the estimation of covariance parameters

The previous equations were at given $\boldsymbol{\mu}$ and k. In practice, however, trend and/or covariance parameters often have to be estimated. Let us consider the case of known μ and k that depends on a vector of "hyperparameters" ψ.

About the estimation of covariance parameters

The previous equations were at given $\boldsymbol{\mu}$ and k. In practice, however, trend and/or covariance parameters often have to be estimated. Let us consider the case of known μ and k that depends on a vector of "hyperparameters" ψ.

Several approaches do exist for dealing with the unknown ψ: Maximum Likelihood Estimation (MLE), Cross-validation (CV), but also Bayesian approaches involving sampling algorithms such as McMC, SMC, etc.

About the estimation of covariance parameters

The previous equations were at given $\boldsymbol{\mu}$ and k. In practice, however, trend and/or covariance parameters often have to be estimated. Let us consider the case of known μ and k that depends on a vector of "hyperparameters" ψ.

Several approaches do exist for dealing with the unknown ψ: Maximum Likelihood Estimation (MLE), Cross-validation (CV), but also Bayesian approaches involving sampling algorithms such as McMC, SMC, etc.

Let us present a brief overview of the MLE approach, probably the most implemented (although not necessarily the most robust) option.

A brief overview of MLE (back to Branin)

Let us denote by $K(\psi)$ the covariance matrix of responses, say $k(\boldsymbol{X}_n, \boldsymbol{X}_n; \psi)$, under the assumption of covariance hyperparameters with value ψ.

A brief overview of MLE (back to Branin)

Let us denote by $K(\psi)$ the covariance matrix of responses, say $k(\boldsymbol{X}_n, \boldsymbol{X}_n; \psi)$, under the assumption of covariance hyperparameters with value ψ.

The principle of MLE is to search for a value of ψ under which it would have been the most likely to observe the responses $\boldsymbol{z}_n$.

Under GP model assumptions, $\boldsymbol{Z}_{\boldsymbol{X}_n} \sim \mathcal{N}(\mu(\boldsymbol{X}_n), K(\psi))$. The likelihood then writes as the probability density of $\boldsymbol{Z}_{\boldsymbol{X}_n}$ at point $\boldsymbol{z}_n$, seen as a function of ψ:

$$L(\psi; \boldsymbol{z}_n) = (2\pi)^{-n/2} \det(K(\psi))^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{z}_n - \mu(\boldsymbol{X}_n))' K(\psi)^{-1} (\boldsymbol{z}_n - \mu(\boldsymbol{X}_n))\right)$$

A brief overview of MLE (back to Branin)

Let us denote by $K(\psi)$ the covariance matrix of responses, say $k(\boldsymbol{X}_n, \boldsymbol{X}_n; \psi)$, under the assumption of covariance hyperparameters with value ψ.

The principle of MLE is to search for a value of ψ under which it would have been the most likely to observe the responses $\boldsymbol{z}_n$.

Under GP model assumptions, $\boldsymbol{Z}_{\boldsymbol{X}_n} \sim \mathcal{N}(\mu(\boldsymbol{X}_n), K(\psi))$. The likelihood then writes as the probability density of $\boldsymbol{Z}_{\boldsymbol{X}_n}$ at point $\boldsymbol{z}_n$, seen as a function of ψ:

$$L(\psi; \boldsymbol{z}_n) = (2\pi)^{-n/2} \det(K(\psi))^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{z}_n - \mu(\boldsymbol{X}_n))' K(\psi)^{-1} (\boldsymbol{z}_n - \mu(\boldsymbol{X}_n))\right)$$

Solving MLE is typically addressed by equivalently minimizing the function

$$\ell(\psi; \boldsymbol{z}_n) = \log(\det(K(\psi))) + (\boldsymbol{z}_n - \mu(\boldsymbol{X}_n))' K(\psi)^{-1} (\boldsymbol{z}_n - \mu(\boldsymbol{X}_n)).$$

A brief overview of MLE

Minimizing ℓ is usually analytically intractable, and numerical optimization algorithms are employed.

A brief overview of MLE

Minimizing ℓ is usually analytically intractable, and numerical optimization algorithms are employed. An elegant trick exists to estimate $\sigma^2 \in (0, +\infty)$ in case k writes as $\sigma^2 \times r$ where r is a given kernel depending on parameters θ.

A brief overview of MLE

Minimizing ℓ is usually analytically intractable, and numerical optimization algorithms are employed. An elegant trick exists to estimate $\sigma^2 \in (0, +\infty)$ in case k writes as $\sigma^2 \times r$ where r is a given kernel depending on parameters θ.

Writing $K(\psi) = \sigma^2 R(\theta)$ where $\psi = (\sigma^2, \theta)$, one can derive the "optimal" σ^2 as a function of θ. A swift calculation leads indeed to

$$\sigma^{2\star}(\theta) = \frac{1}{n}(\boldsymbol{z}_n - \boldsymbol{\mu}(\boldsymbol{X}_n))' R(\theta)^{-1} (\boldsymbol{z}_n - \boldsymbol{\mu}(\boldsymbol{X}_n)).$$

A brief overview of MLE

Minimizing ℓ is usually analytically intractable, and numerical optimization algorithms are employed. An elegant trick exists to estimate $\sigma^2 \in (0, +\infty)$ in case k writes as $\sigma^2 \times r$ where r is a given kernel depending on parameters θ.

Writing $K(\psi) = \sigma^2 R(\theta)$ where $\psi = (\sigma^2, \theta)$, one can derive the "optimal" σ^2 as a function of θ. A swift calculation leads indeed to

$$\sigma^{2\star}(\theta) = \frac{1}{n}(\boldsymbol{z}_n - \boldsymbol{\mu}(\boldsymbol{X}_n))' R(\theta)^{-1} (\boldsymbol{z}_n - \boldsymbol{\mu}(\boldsymbol{X}_n)).$$

Re-injecting the latter equation into ℓ, MLE then boils down to minimizing a function depending solely on θ, the so-called profile (or "concentrated") ℓ:

$$\ell_p(\theta; \boldsymbol{z}_n) = \log(\det(\sigma^{2\star}(\theta) R(\theta)))$$

Towards Universal Kriging

Another situation where an elegant concentration of ℓ is feasible is when k depends on ψ and $\boldsymbol{\mu}$ linearly depends on p basis functions $f_1, \ldots, f_p$:

$$\mu(\mathbf{x}) = \sum_{i=1}^{p} \beta_i f_i(\mathbf{x}),$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is assumed unknown.

Towards Universal Kriging

Another situation where an elegant concentration of ℓ is feasible is when k depends on ψ and $\boldsymbol{\mu}$ linearly depends on p basis functions $f_1, \ldots, f_p$:

$$\mu(\mathbf{x}) = \sum_{i=1}^{p} \beta_i f_i(\mathbf{x}),$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is assumed unknown. Then, setting $F = (f_j(\mathbf{x}_i))_{1 \leq i \leq n, 1 \leq j \leq p}$, we have $\mu(\boldsymbol{X}_n) = F\boldsymbol{\beta}$, and maximizing the likelihood with respect to $\boldsymbol{\beta}$ at fixed covariance parameters (say ψ again) leads to:

$$\boldsymbol{\beta}^\star(\psi) = (F' K(\psi)^{-1} F)^{-1} F' K(\psi)^{-1} \boldsymbol{z}_n.$$

Towards Universal Kriging

Another situation where an elegant concentration of ℓ is feasible is when k depends on ψ and $\boldsymbol{\mu}$ linearly depends on p basis functions $f_1, \ldots, f_p$:

$$\mu(\mathbf{x}) = \sum_{i=1}^{p} \beta_i f_i(\mathbf{x}),$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is assumed unknown. Then, setting $F = (f_j(\mathbf{x}_i))_{1 \leq i \leq n, 1 \leq j \leq p}$, we have $\mu(\boldsymbol{X}_n) = F\boldsymbol{\beta}$, and maximizing the likelihood with respect to $\boldsymbol{\beta}$ at fixed covariance parameters (say ψ again) leads to:

$$\boldsymbol{\beta}^\star(\psi) = (F' K(\psi)^{-1} F)^{-1} F' K(\psi)^{-1} \boldsymbol{z}_n.$$

Plugging-in $\boldsymbol{\beta}^\star(\psi)$ in the predictor and inflating the conditional (co)variance accordingly leads to the "Universal Kriging" equations (See also particular case of "Ordinary Kriging", where $p = 1$ and μ is a constant; Eqs).

NB: In a Bayesian set-up where an improper uniform prior is put on $\boldsymbol{\beta}$, one even recovers a GP posterior distribution.

Selected references

M.L. Stein (1999).
Interpolation of Spatial Data, Some Theory for Kriging.
Springer

R. Adler and J. Taylor (2007).
Random Fields and Geometry.
Springer

M. Scheuerer (2009).
A Comparison of Models and Methods for Spatial Interpolation in Statistics and Numerical Analysis.
PhD thesis of Georg-August Universität Göttingen

O. Roustant, D. Ginsbourger, Y. Deville (2012).
DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization.
Journal of Statistical Software, 51(1), 1-55.

M. Schlather, A. Malinowski, P. J. Menck, M. Oesting and K. Strokorb (2015).
Analysis, Simulation and Prediction of Multivariate Random Fields with Package RandomFields.
Journal of Statistical Software, 63, 8, 1-25.

Selected references

B. Rajput and S. Cambanis (1972).

Gaussian processes and Gaussian measures.

Ann. Math. Statist. 43 (6), 1944-1952.

A. O'Hagan (1978).

Curve fitting and optimal design for prediction.

Journal of the Royal Statistical Society, Series B, 40(1):1-42.

H. Omre and K. Halvorsen (1989).

The bayesian bridge between simple and universal kriging.

Mathematical Geology, 22 (7):767-786.

M. S. Handcock and M. L. Stein (1993).

A bayesian analysis of kriging.

Technometrics, 35(4):403-410.

A.W. Van der Vaart and J. H. Van Zanten (2008).

Rates of contraction of posterior distributions based on Gaussian process priors.

Annals of Statistics, 36:1435-1463.

Outline

Some examples of p.d. kernels and GPs

Let us start by a very classical example (for $d = 1$): the Brownian motion $W = (W_t)_{t \in D}$ over $D = [0, +\infty)$. Let us define W (in distribution) as follows:

- $W_0 = 0$,
- for any $t \in D$ and $h > 0$, $W_{t+h} - W_t \sim \mathcal{N}(0, h)$,
- and for any $t_1, t_2, t_3, t_4 \in D$ with $t_1 \leq t_2 \leq t_3 \leq t_4$, the increments $W_{t_4} - W_{t_3}$ and $W_{t_2} - W_{t_1}$ are independent.

Some examples of p.d. kernels and GPs

Let us start by a very classical example (for $d = 1$): the Brownian motion $W = (W_t)_{t \in D}$ over $D = [0, +\infty)$. Let us define W (in distribution) as follows:

- $W_0 = 0$,
- for any $t \in D$ and $h > 0$, $W_{t+h} - W_t \sim \mathcal{N}(0, h)$,
- and for any $t_1, t_2, t_3, t_4 \in D$ with $t_1 \leq t_2 \leq t_3 \leq t_4$, the increments $W_{t_4} - W_{t_3}$ and $W_{t_2} - W_{t_1}$ are independent.

Such conditions define a GP; there remains to work out its expectation and covariance functions. First, for $t \in D$ the two first conditions imply that

$$m(t) = \mathbb{E}[W_t] = \mathbb{E}[W_0 + W_t - W_0] = \mathbb{E}[W_0] + \mathbb{E}[W_t - W_0] = 0 + 0 = 0.$$

Some examples of p.d. kernels and GPs

Let us start by a very classical example (for $d = 1$): the Brownian motion $W = (W_t)_{t \in D}$ over $D = [0, +\infty)$. Let us define W (in distribution) as follows:

- $W_0 = 0$,
- for any $t \in D$ and $h > 0$, $W_{t+h} - W_t \sim \mathcal{N}(0, h)$,
- and for any $t_1, t_2, t_3, t_4 \in D$ with $t_1 \leq t_2 \leq t_3 \leq t_4$, the increments $W_{t_4} - W_{t_3}$ and $W_{t_2} - W_{t_1}$ are independent.

Such conditions define a GP; there remains to work out its expectation and covariance functions. First, for $t \in D$ the two first conditions imply that

$$m(t) = \mathbb{E}[W_t] = \mathbb{E}[W_0 + W_t - W_0] = \mathbb{E}[W_0] + \mathbb{E}[W_t - W_0] = 0 + 0 = 0.$$

Second, taking two points $t, t' \in D$ (assuming, say, that $t < t'$), the third condition implies that $W_{t'} - W_t$ is independent of $W_t - W_0$. Consequently,

$$\begin{aligned} k_{BM}(t, t') &= \mathbb{E}[W_t W_{t'}] = \mathbb{E}[(W_t - W_0)(W_t - W_0 + W_{t'} - W_t)] \\ &= \mathbb{E}[(W_t - W_0)^2] + \mathbb{E}[(W_t - W_0)(W_{t'} - W_t)] = t + 0 = t = \min(t, t'). \end{aligned}$$

Examples of covariance kernels and GPs (cont'd)

Another famous covariance function stems from the so-called "Brownian Bridge" (ending in 0) $B = (B_t)_{t\in[0,1]}$. Let us first restrict W to $D = [0, 1]$, obtaining a centred process with covariance $k(t, t') = \min(t, t')$ over $[0, 1]^2$.

The distribution of B is then obtained by conditioning W on $W_1 = 0$, thus obtaining the mean $m_B(t) = 0$ and covariance kernel

$$k_{BB}(t, t') = \min(t, t') - tt' = \min(t, t')(1 - \max(t, t')).$$

Examples of covariance kernels and GPs (cont'd)

Another famous covariance function stems from the so-called "Brownian Bridge" (ending in 0) $B = (B_t)_{t\in[0,1]}$. Let us first restrict W to $D = [0, 1]$, obtaining a centred process with covariance $k(t, t') = \min(t, t')$ over $[0, 1]^2$.

The distribution of B is then obtained by conditioning W on $W_1 = 0$, thus obtaining the mean $m_B(t) = 0$ and covariance kernel

$$k_{BB}(t, t') = \min(t, t') - tt' = \min(t, t')(1 - \max(t, t')).$$

Another covariance function of interest can be obtained by integrating W. Defining $(I_t)_{t\in D}$ (with $D = [0, +\infty)$ again) by $I_t = \int_0^t B_u \mathrm{d}u$, we obtain a new centred GP with covariance

$$\begin{aligned} k_{IBM}(t, t') &= \int_0^t \int_0^{t'} \min(u, v)\mathrm{d}u\mathrm{d}v \\ &= \min(t, t')^3/3 + (\max(t, t') - \min(t, t'))\min(t, t')^2/2. \end{aligned}$$

Examples of covariance kernels and GPs (cont'd)

Without entering into much detail, let us list a few further examples of 1-dimensional GPs / associated covariance kernels:

- For $D = [0, 1]$ and $H \in (0, 1)$, $k_{fBM}(t, t') = \frac{1}{2}(|t|^{2H} + |t'|^{2H} - |t - t'|^{2H})$ is the covariance kernel of the *fractional (or "fractal") Brownian Motion* with Hurst coefficient H,
- $k_{\text{triang}}(t, t') = (1 - |t - t'|)^+$ is the "triangular" kernel over $D = \mathbb{R}$,
- Defining $Z_t = \zeta_1 \cos(\omega t) + \zeta_2 \sin(\omega t)$, where $\zeta_1, \zeta_2 \sim \mathcal{N}(0, \sigma^2)$ independently ($\sigma > 0$) and $\omega > 0$, one obtains $k(t, t') = \cos(\omega(t' - t))$,
- $k_{OU}(t, t') = e^{-|t-t'|}$ is called exponential kernel and characterizes the *Ornstein-Uhlenbeck process*.
- $k(t, t') = e^{-|t-t'|^2}$ is the squared-exponential kernel.

Examples of covariance kernels and GPs (cont'd)

Previous k's from real-valued one-dimensional settings can be generalize in a number of ways. Let us review a few simple examples.

- One obtains an admissible k on $[0,+\infty)^d \times [0,+\infty)^d$ by taking $k(\mathbf{x},\mathbf{x}') = \prod_{i=1}^d \min(x_i, x_i')$ where the $x_i^{(')}$'s are the coordinates of $\mathbf{x}^{(')}$. The associated centred GP over $[0,+\infty)^d$ is called "Brownian Sheet".

- The exponential and Gaussian kernels can be generalized to $\mathbb{R}^d \times \mathbb{R}^d$ by taking $k(\mathbf{x},\mathbf{x}') = \exp(-||\mathbf{x}-\mathbf{x}'||)$ and $k(\mathbf{x},\mathbf{x}') = \exp(-||\mathbf{x}-\mathbf{x}'||^2)$, respectively, where $||\cdot||$ refers to the Euclidean norm on $\mathbb{R}^d$.

- From a different perspective, one can define a particular complex-valued GP by taking $Z_{\mathbf{x}} = \zeta \exp^{-\mathrm{i}\langle \mathbf{x},\boldsymbol{\omega}\rangle}$ where $\zeta \sim \mathcal{N}(0,\sigma^2)$ ($\sigma > 0$) and $\boldsymbol{\omega} \in \mathbb{R}^d$. Such Z is centred and has (complex) covariance

$$k(\mathbf{x},\mathbf{x}') = \mathrm{Cov}(Z_x, Z_x') = \mathbb{E}[Z_{\mathbf{x}}\overline{Z_{\mathbf{x}'}}] = \sigma^2 \exp^{-\mathrm{i}\langle \mathbf{x},\boldsymbol{\omega}\rangle} \exp^{\mathrm{i}\langle \mathbf{x}',\boldsymbol{\omega}\rangle} = \exp^{-\mathrm{i}\langle \mathbf{x}-\mathbf{x}',\boldsymbol{\omega}\rangle}.$$

A necessary and sufficient condition of admissibility

A common point about all kernels reviewed so far is that, for ad hoc D, if one takes any $n \geq 1$ and arbitrary points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and complex numbers $a_1, \ldots, a_n \in \mathbb{C}$, the following holds:

$$0 \leq \mathrm{Var}\left[\sum_{i=1}^{n} a_i Z_{\mathbf{x}_i}\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \overline{a_j} k(\mathbf{x}_i, \mathbf{x}_j).$$

A necessary and sufficient condition of admissibility

A common point about all kernels reviewed so far is that, for ad hoc D, if one takes any $n \geq 1$ and arbitrary points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and complex numbers $a_1, \ldots, a_n \in \mathbb{C}$, the following holds:

$$0 \leq \mathrm{Var}\left[\sum_{i=1}^{n} a_i Z_{\mathbf{x}_i}\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \overline{a_j} k(\mathbf{x}_i, \mathbf{x}_j).$$

This property is indeed necessary for k to be an admissible covariance. Furthermore, it turns out that any k possessing this property is a covariance kernel (there exists some (Gaussian) random process with this k).

From p.d. kernels to function approximation

For an introduction to the mathematical foundations of p.d. kernels and their use in function approximation, see notably the following references:


C. Berg, J. P. R. Christensen and P. Ressel (1984)
Harmonic Analysis on Semigroups. Theory of Positive Definite and Related Functions
Springer-Verlag


A. Berlinet, C. Thomas-Agnan (2004)
Reproducing Kernel Hilbert Spaces in Probability and Statistics
Kluwer Academic Publishers

Choosing p.d. kernels?

In practice, choosing an adapted k for an objective f (about which limited information may be available) is both a crucial and difficult task.

Choosing p.d. kernels?

In practice, choosing an adapted k for an objective f (about which limited information may be available) is both a crucial and difficult task.

Typically, k is chosen among some well-known p.d. kernel families, often among "shift-invariant" (a.k.a. "stationary") kernels, i.e. functions of $\mathbf{x} - \mathbf{x}'$.

Examples: Generalized Exponential (including Gaussian) kernels, Matérn kernels, and more generally kernels obtained via the Bochner theorem.

Bochner theorem

By a slight abuse of notation, we denote stationary kernels on $D = \mathbb{R}^d$ ($d \geq 1$) by $k : \mathbf{h} \in \mathbb{R}^d \longrightarrow k(\mathbf{h}) \in \mathbb{C}$.

Theorem (Bochner's theorem)

A continuous $k : \mathbf{h} \in \mathbb{R}^d \longrightarrow k(\mathbf{h}) \in \mathbb{C}$ is positive definite if and only if it is the Fourier transform of a finite non-negative Borel measure ν on $\mathbb{R}^d$, i.e.

$$k(\mathbf{h}) = \hat{\nu}(\mathbf{h}) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{h}, \boldsymbol{\omega} \rangle} d\nu(\boldsymbol{\omega})$$

See for instance Wendland 2005 (Chap. 6) for a proof.

Bochner theorem

By a slight abuse of notation, we denote stationary kernels on $D = \mathbb{R}^d$ ($d \geq 1$) by $k : \mathbf{h} \in \mathbb{R}^d \longrightarrow k(\mathbf{h}) \in \mathbb{C}$.

Theorem (Bochner's theorem)

A continuous $k : \mathbf{h} \in \mathbb{R}^d \longrightarrow k(\mathbf{h}) \in \mathbb{C}$ is positive definite if and only if it is the Fourier transform of a finite non-negative Borel measure ν on $\mathbb{R}^d$, i.e.

$$k(\mathbf{h}) = \hat{\nu}(\mathbf{h}) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{h}, \boldsymbol{\omega} \rangle} d\nu(\boldsymbol{\omega})$$

See for instance Wendland 2005 (Chap. 6) for a proof.

By playing on the "spectral measure" ν one can hence generate all continuous stationary p.d. kernels on $\mathbb{R}^d$. For the case of a measure ν admitting a density $q(\boldsymbol{\omega}) = \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}(\boldsymbol{\omega})$ w.r.t. the Lebesgue measure λ, k is hence characterized by its *spectral density* q.

A few 1-dimensional examples

- Triangular: $k(h) := c(a - |h|)^+$ $\left(q(\omega) \sim \frac{c(1-\cos(a\omega))}{\pi\omega^2}\right)$
- Matérn $\nu = 3/2$: $k(h) \sim \alpha^{-3} e^{-\alpha|t|}(1 + \alpha|t|)$ $(q(\omega) \sim (\alpha^2 + \omega^2)^{-2})$
- Gauss: $k(h) \sim e^{-(\frac{t}{\theta})^2}$ $(q(\omega) \sim e^{-\theta^2\omega^2})$

M. Stein (Springer, 1999)

Interpolation of Spatial Data. Some Theory for Kriging

More on spectral densities of Matérn kernels ($d \geq 1$)

Matérn kernels can be characterized using the Hancock and Wallis parametrization (1994) mentioned in Stein (1999) (here $\sigma = 1$):

$$q(\omega) = \frac{c(\nu, \rho)}{\left(\frac{4\nu}{\rho^2} + ||\omega||^2\right)^{\nu + d/2}}$$

where $c(\nu, \rho) = \frac{\Gamma\left(\nu + \frac{d}{2}\right)(4\nu)^\nu}{\pi^{d/2}\Gamma(\nu)\rho^{2\nu}}$.

More on spectral densities of Matérn kernels ($d \geq 1$)

Matérn kernels can be characterized using the Hancock and Wallis parametrization (1994) mentioned in Stein (1999) (here $\sigma = 1$):

$$q(\boldsymbol{\omega}) = \frac{c(\nu, \rho)}{\left(\frac{4\nu}{\rho^2} + ||\boldsymbol{\omega}||^2\right)^{\nu + d/2}}$$

where $c(\nu, \rho) = \frac{\Gamma\left(\nu + \frac{d}{2}\right)(4\nu)^\nu}{\pi^{d/2}\Gamma(\nu)\rho^{2\nu}}$. The corresponding ("*isotropic*") p.d. kernel is

$$k(\mathbf{h}) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}||\mathbf{h}||}{\rho}\right)^\nu \mathcal{K}_\nu \left(\frac{2\nu^{1/2}||\mathbf{h}||}{\rho}\right)$$

where $\mathcal{K}_\nu$ is a *modified Bessel function of the third kind*. More tractable expressions can be obtained for $\nu = \frac{1}{2}, \frac{3}{2}, \frac{5}{2} \ldots$

More on spectral densities of Matérn kernels ($d \geq 1$)

Matérn kernels can be characterized using the Hancock and Wallis parametrization (1994) mentioned in Stein (1999) (here $\sigma = 1$):

$$q(\boldsymbol{\omega}) = \frac{c(\nu, \rho)}{\left(\frac{4\nu}{\rho^2} + ||\boldsymbol{\omega}||^2\right)^{\nu + d/2}}$$

where $c(\nu, \rho) = \frac{\Gamma\left(\nu + \frac{d}{2}\right)(4\nu)^\nu}{\pi^{d/2}\Gamma(\nu)\rho^{2\nu}}$. The corresponding ("*isotropic*") p.d. kernel is

$$k(\mathbf{h}) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}||\mathbf{h}||}{\rho}\right)^\nu \mathcal{K}_\nu \left(\frac{2\nu^{1/2}||\mathbf{h}||}{\rho}\right)$$

where $\mathcal{K}_\nu$ is a *modified Bessel function of the third kind*. More tractable expressions can be obtained for $\nu = \frac{1}{2}, \frac{3}{2}, \frac{5}{2} \ldots$ See Stein (1999) for more on this class and Wendland (2005) –chap. 5– for more on Bessel functions.

More on isotropic p.d. kernels in $\mathbb{R}^d$

The Matérn (class of) kernels considered previously are one among many *isotropic p.d. kernels on* $\mathbb{R}^d$, i.e. p.d. kernels that write as

$$k(\mathbf{x}, \mathbf{x}') = \kappa(r)$$

where $r = ||\mathbf{x} - \mathbf{x}'||_{\mathbb{R}^d}$, and $\kappa : \mathbb{R}_+ \longrightarrow \mathbb{R}$ is also often (by a slight abusive of language) referred to as positive definite. Such k's are also called *radial*.

More on isotropic p.d. kernels in $\mathbb{R}^d$

The Matérn (class of) kernels considered previously are one among many *isotropic p.d. kernels on* $\mathbb{R}^d$, i.e. p.d. kernels that write as

$$k(\mathbf{x}, \mathbf{x}') = \kappa(r)$$

where $r = ||\mathbf{x} - \mathbf{x}'||_{\mathbb{R}^d}$, and $\kappa : \mathbb{R}_+ \longrightarrow \mathbb{R}$ is also often (by a slight abusive of language) referred to as positive definite. Such k's are also called *radial*.

Definition (Cf. Wendland 2005): A function $\Phi : \mathbb{R}^d \longrightarrow \mathbb{R}$ is said to be *radial* if there exists $\phi : [0, +\infty) \longrightarrow \mathbb{R}$ such that $\Phi(\mathbf{h}) = \phi(||\mathbf{h}||_2)$ for all $\mathbf{h} \in \mathbb{R}^d$.

A number κ leading to radial p.d. kernels in $\mathbb{R}^d$ do exist and have been studied by generations of mathematicians. Some depend on d, some do not!

More on isotropic p.d. kernels in $\mathbb{R}^d$

Let us review of few examples.

- $\kappa(r) = e^{-r^p}$ $(0 < p \leq 2)$ "Generalized exponential"

More on isotropic p.d. kernels in $\mathbb{R}^d$

Let us review of few examples.

- $\kappa(r) = e^{-r^p}$ $(0 < p \leq 2)$ "Generalized exponential"
- $\kappa(r) = (c^2 + r^2)^{-\beta}$ $(c, \beta > 0)$ "Inverse multiquadrics"

More on isotropic p.d. kernels in $\mathbb{R}^d$

Let us review of few examples.

- $\kappa(r) = e^{-r^p}$ $(0 < p \leq 2)$ "Generalized exponential"
- $\kappa(r) = (c^2 + r^2)^{-\beta}$ $(c, \beta > 0)$ "Inverse multiquadrics"
- $\kappa(r) = (1 - r)_+^\ell$ where $(x)_+ = \max(0, x)$ "Truncated power kernel")

More on isotropic p.d. kernels in $\mathbb{R}^d$

Let us review of few examples.

- $\kappa(r) = e^{-r^p}$ $(0 < p \leq 2)$ "Generalized exponential"
- $\kappa(r) = (c^2 + r^2)^{-\beta}$ $(c, \beta > 0)$ "Inverse multiquadrics"
- $\kappa(r) = (1 - r)_+^\ell$ where $(x)_+ = \max(0, x)$ "Truncated power kernel")

While the two first kernels are (strictly) positive definite for all $d \geq 1$, for the third one one needs to restrict to $\ell \geq \lfloor d/2 \rfloor + 1$ to get this property.

More on isotropic p.d. kernels in $\mathbb{R}^d$

Let us review of few examples.

- $\kappa(r) = e^{-r^p}$ $(0 < p \leq 2)$ "Generalized exponential"
- $\kappa(r) = (c^2 + r^2)^{-\beta}$ $(c, \beta > 0)$ "Inverse multiquadrics"
- $\kappa(r) = (1 - r)_+^\ell$ where $(x)_+ = \max(0, x)$ "Truncated power kernel")

While the two first kernels are (strictly) positive definite for all $d \geq 1$, for the third one one needs to restrict to $\ell \geq \lfloor d/2 \rfloor + 1$ to get this property.

Is it possible to characterize radial p.d. functions defined in terms of one κ valid in any dimension?

More on isotropic p.d. kernels in $\mathbb{R}^d$

Let us review of few examples.

- $\kappa(r) = e^{-r^p}$ $(0 < p \leq 2)$ "Generalized exponential"
- $\kappa(r) = (c^2 + r^2)^{-\beta}$ $(c, \beta > 0)$ "Inverse multiquadrics"
- $\kappa(r) = (1 - r)_+^\ell$ where $(x)_+ = \max(0, x)$ "Truncated power kernel")

While the two first kernels are (strictly) positive definite for all $d \geq 1$, for the third one one needs to restrict to $\ell \geq \lfloor d/2 \rfloor + 1$ to get this property.

Is it possible to characterize radial p.d. functions defined in terms of one κ valid in any dimension? Yes, thanks to completely monotone functions!

More on isotropic p.d. kernels in $\mathbb{R}^d$

Definition (Cf. Wendland 2005): A function ϕ is called completely monotone on $(0, +\infty)$ if it satisfies $\phi \in C^\infty(0, +\infty)$ and

$$(-1)^\ell \phi^{(\ell)}(r) \geq 0$$

for all $\ell \in \mathbb{N}$ and all $r > 0$. The function ϕ is called completely monotone on $[0, +\infty)$ if it is in addition in $C[0, +\infty)$.

More on isotropic p.d. kernels in $\mathbb{R}^d$

Definition (Cf. Wendland 2005): A function ϕ is called completely monotone on $(0, +\infty)$ if it satisfies $\phi \in C^\infty(0, +\infty)$ and

$$(-1)^\ell \phi^{(\ell)}(r) \geq 0$$

for all $\ell \in \mathbb{N}$ and all $r > 0$. The function ϕ is called completely monotone on $[0, +\infty)$ if it is in addition in $C[0, +\infty)$.

Theorem (Schoenberg, Cf. Wendland 2005)

A function $\phi : [0, +\infty) \longrightarrow \mathbb{R}$ is completely monotone on $[0, +\infty)$ if and only if $\Phi := \phi(|| \cdot ||_2^2)$ is positive definite on every $\mathbb{R}^d$.

More on isotropic p.d. kernels in $\mathbb{R}^d$

Definition (Cf. Wendland 2005): A function ϕ is called completely monotone on $(0, +\infty)$ if it satisfies $\phi \in C^\infty(0, +\infty)$ and

$$(-1)^\ell \phi^{(\ell)}(r) \geq 0$$

for all $\ell \in \mathbb{N}$ and all $r > 0$. The function ϕ is called completely monotone on $[0, +\infty)$ if it is in addition in $C[0, +\infty)$.

Theorem (Schoenberg, Cf. Wendland 2005)

A function $\phi : [0, +\infty) \longrightarrow \mathbb{R}$ is completely monotone on $[0, +\infty)$ if and only if $\Phi := \phi(|| \cdot ||_2^2)$ is positive definite on every $\mathbb{R}^d$.

Application: the inverse multiquadrics is p.d. in any dim. for $c, \beta > 0$.

Nota Bene: geometric anisotropy

Starting from any isotropic p.d. kernel, it is always possible to generalize it and obtain *(geometric) anisotropic* p.d. kernels through orthogonal transformations and dilatations, by defining

$$k(\mathbf{x}, \mathbf{x}') = \kappa\left((\mathbf{x} - \mathbf{x}')^T \Sigma (\mathbf{x} - \mathbf{x}')\right)$$

where Σ is a real-valued symmetric (strictly!) positive definite matrix.

Other ways of defining p.d. kernels: overview

Kernels that write as functions of $\langle \mathbf{x}, \mathbf{x}' \rangle$ (as the previously presented radial p.d. kernels on the sphere) are also called *zonal kernels* in G. E. Fasshauer's review paper below, were examples of zonal kernels are discussed:

Fasshauer, G. E. (2011)
Positive definite kernels: past, present and future
Dolomites Research Notes on Approximation, 4:21-63

The following paper also includes alternative classes of p.d. kernels:

T. Hofmann, B. Schölkopf, A.J. Smola (2008)
Kernel methods in machine learning
The Annals of Statistics, Vol. 36, No. 3, 1171-1220.

Other ways of defining p.d. kernels: overview

Kernels that write as functions of $\langle \mathbf{x}, \mathbf{x}' \rangle$ (as the previously presented radial p.d. kernels on the sphere) are also called *zonal kernels* in G. E. Fasshauer's review paper below, were examples of zonal kernels are discussed:

Fasshauer, G. E. (2011)
Positive definite kernels: past, present and future
Dolomites Research Notes on Approximation, 4:21-63

The following paper also includes alternative classes of p.d. kernels:

T. Hofmann, B. Schölkopf, A.J. Smola (2008)
Kernel methods in machine learning
The Annals of Statistics, Vol. 36, No. 3, 1171-1220.

Overall, the notion of **scalar product** plays a crucial role in p.d. kernels.

Other ways of defining p.d. kernels: Mercer theorem

For continuous p.d. kernels –say real-valued, defined on a compact set $D \subset \mathbb{R}^d$– a fruitful approach is to consider the following operator T_k on $L^2(D)$:

$$g \longrightarrow T_k(g)(\cdot) = \int_D g(\mathbf{x}')k(\cdot, \mathbf{x}')\mathrm{d}\lambda(\mathbf{x}')$$

where λ refers to the Lebesgue measure (generalizations do exist) on $\mathbb{R}^d$.

Under our continuity and compactness conditions on T_k there exist $(\varphi_j(\cdot))_{j\in\mathbb{N}^*}$ forming an orthonormal system of $L^2(D)$ and $(\lambda_j)_{j\in\mathbb{N}^*}$ non-negative such that

$$\forall j \in \mathbb{N}\ T_k(\varphi_j) = \lambda_j\varphi_j$$

and this leads to the Mercer decomposition (1909):

$$k(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \lambda_j\varphi_j(\mathbf{x})\varphi_j(\mathbf{x}').$$

See Adler & Taylor, Steinwart and more for detail on the convergence, etc.

Basic principle of the Karhunen-Loève expansion

Assuming D compact and k continuous, the Mercer theorem ensures the existence of an orthonormal basis $(\varphi_j)_{j\geq 1}$ of $L^2(D)$ such that

$$k(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{+\infty} \lambda_j \varphi_j(\mathbf{x}) \varphi_j(\mathbf{x}')$$

Basic principle of the Karhunen-Loève expansion

Assuming D compact and k continuous, the Mercer theorem ensures the existence of an orthonormal basis $(\varphi_j)_{j\geq 1}$ of $L^2(D)$ such that

$$k(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{+\infty} \lambda_j \varphi_j(\mathbf{x}) \varphi_j(\mathbf{x}')$$

The KL expansion of a GRF Z then consists in representing it under the form

$$Z_{\mathbf{x}} = \sum_{j=1}^{+\infty} \sqrt{\lambda_j} \zeta_j \varphi_j(\mathbf{x})$$

where the ζ_j's are i.i.d. standard Gaussian random variables.

Deriving the eigenfunctions: a Fredholm problem

Given a GRF Z of covariance kernel k, finding the basis functions φ_j ($j \geq 1$) is the key to the KL decomposition of Z.

This is done by solving the following integral equation:

$$\int_D k(\mathbf{x}, \mathbf{x}')g(\mathbf{x})d\mu(\mathbf{x}) = \lambda g(\mathbf{x}'),$$

called a *Fredholm problem*.

Deriving the eigenfunctions: a Fredholm problem

Given a GRF Z of covariance kernel k, finding the basis functions φ_j ($j \geq 1$) is the key to the KL decomposition of Z.

This is done by solving the following integral equation:

$$\int_D k(\mathbf{x}, \mathbf{x}')g(\mathbf{x})d\mu(\mathbf{x}) = \lambda g(\mathbf{x}'),$$

called a *Fredholm problem*.

When possible, the latter is solved analytically by using calculus.

Example: KL expansion of the Brownian Motion

For the covariance kernel of the BM, $k(t, t') = \min(t, t')$, the eigenvalues and eigenfunctions are solutions to the following Fredholm problem:

$$\int_0^1 \min(t, t')\varphi(t)dt = \lambda\varphi(t')$$

Example: KL expansion of the Brownian Motion

For the covariance kernel of the BM, $k(t, t') = \min(t, t')$, the eigenvalues and eigenfunctions are solutions to the following Fredholm problem:

$$\int_0^1 \min(t, t')\varphi(t)dt = \lambda\varphi(t')$$

It can be shown by solving a differential equation that the solutions are

$$\lambda_j = \frac{1}{\pi^2(j - \frac{1}{2})^2}$$

$$\varphi_j(t) = \sqrt{2}\sin\left(\left(j - \frac{1}{2}\right) \times \pi t\right)$$

R.J. Adler and J.E. Taylor (Springer, 2007)

Random Fields and Geometry

Example: KL expansion of the Brownian Motion

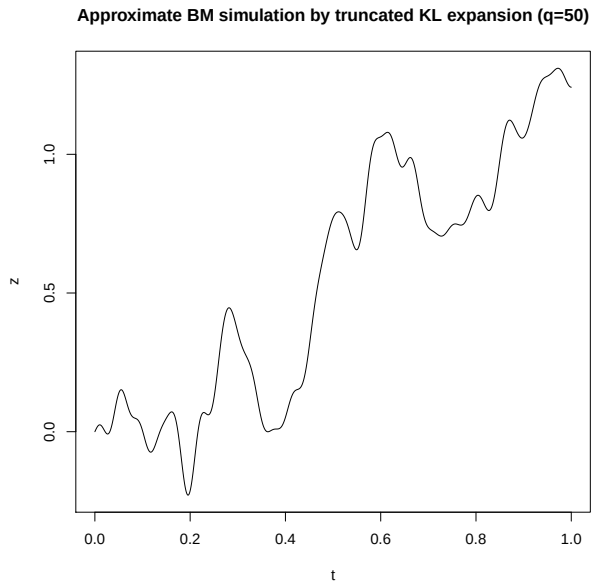
Let us simulate the Brownian Motion using a truncated KL expansion:

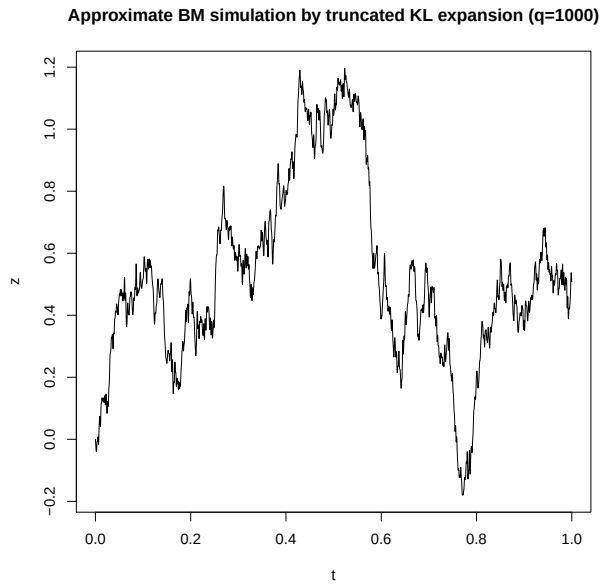
```
m <- 10000
t <- seq(0,1,,m)
v <- function(t,k){sqrt(2)*sin((k-0.5)*pi*t)}
lambda <- function(k){1/(pi*(k-0.5))^2}

q <- 1000
KL <- rep(0,m)
for(i in seq(1,q)){
KL <- KL + sqrt(lambda(i))*rnorm(1)*v(t,i)}
```

Example: KL expansion of the Brownian Motion

Here are two simulation results based on the truncated KL expansion of the Brownian Motion, respectively with $q = 50$ and $q = 1000$:





The simulations are not exact, but can be performed at a continuous set. The ζ_j's can be stored, and the corresponding path evaluated at a new point later.

A few selected references

M.L. Stein (1999).
Interpolation of Spatial Data, Some Theory for Kriging
Springer

M. Scheuerer (2009).
A Comparison of Models and Methods for Spatial Interpolation in Statistics and Numerical Analysis
PhD thesis of Georg-August Universität Göttingen

C. E. Rasmussen and C.K.I. Williams (2006).
Gaussian Processes for Machine Learning
MIT Press

R. Adler and J. Taylor (2007).
Random Fields and Geometry
Springer

I. Steinwart (2017).

Convergence Types and Rates in Generic Karhune–Loève Expansions with Applications to Sample Path Properties
arXiv:1403.1040v3 [math.PR]

Outline

Further properties of p.d. kernels (back)

Further general properties can be derived for p.d. kernels, including:

- Products of p.d. kernels are p.d. kernels
- If $\sigma : D \longrightarrow D$ is a bijection, $k(\mathbf{x}, \mathbf{x}')$ is a p.d. kernel if and only if $k(\sigma(\mathbf{x}), \sigma(\mathbf{x}'))$ is a p.d. kernel
- For all $\mathbf{x}, \mathbf{x}' \in D$ $|k(\mathbf{x}, \mathbf{x}')| \leq \sqrt{k(\mathbf{x}, \mathbf{x})}\sqrt{k(\mathbf{x}', \mathbf{x}')}$
- The function

  $$d_k : (\mathbf{x}, \mathbf{x}') \in D^2 \longrightarrow d_k(\mathbf{x}, \mathbf{x}') = \sqrt{k(\mathbf{x}, \mathbf{x}) + k(\mathbf{x}', \mathbf{x}') - 2\Re(k(\mathbf{x}, \mathbf{x}'))}$$

  defines a (pseudo-)distance on D.

Further properties of p.d. kernels (back)

Further general properties can be derived for p.d. kernels, including:

- Products of p.d. kernels are p.d. kernels
- If $\sigma : D \longrightarrow D$ is a bijection, $k(\mathbf{x}, \mathbf{x}')$ is a p.d. kernel if and only if $k(\sigma(\mathbf{x}), \sigma(\mathbf{x}'))$ is a p.d. kernel
- For all $\mathbf{x}, \mathbf{x}' \in D$ $|k(\mathbf{x}, \mathbf{x}')| \leq \sqrt{k(\mathbf{x}, \mathbf{x})}\sqrt{k(\mathbf{x}', \mathbf{x}')}$
- The function

$$d_k : (\mathbf{x}, \mathbf{x}') \in D^2 \longrightarrow d_k(\mathbf{x}, \mathbf{x}') = \sqrt{k(\mathbf{x}, \mathbf{x}) + k(\mathbf{x}', \mathbf{x}') - 2\Re(k(\mathbf{x}, \mathbf{x}'))}$$

  defines a (pseudo-)distance on D.

Note also that positive definiteness can be generalized as follows: $k : (\mathbf{x}, \mathbf{x}') \in D^2 \longrightarrow \mathbb{C}$ is called *conditionally positive definite* (*c.p.d.*) if it is hermitian and $\sum_{i=1}^{n} \sum_{j=1}^{n} a_i \overline{a_j} k(\mathbf{x}_i, \mathbf{x}_j) \in [0, +\infty)$ for all $n \geq 1$, $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D$ and $a_1, \ldots, a_n \in \mathbb{C}$ s.t $\sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} \mathbf{a_i} = \mathbf{0}$. *C.n.d.* is defined similarly with $(-\infty, 0]$.

RKHS

Reproducing Kernel Hilbert Spaces (RKHS) offer a very convenient framework for function approximation. Here

Definition: A Hilbert space of functions $D \longrightarrow \mathbb{C}$, $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$, is a RKHS if for all $\mathbf{x} \in D$, the evaluation functional $e_{\mathbf{x}} : f \in \mathcal{H} \longrightarrow f(\mathbf{x}) \in \mathbb{C}$ are continuous.

From the so-called *Riesz representation theorem*, for all $\mathbf{x} \in D$ there exists an element of $\mathcal{H}$, denoted here $k_{\mathbf{x}}$, such that $f(\mathbf{x}) = \langle f, k_{\mathbf{x}} \rangle_{\mathcal{H}}$.

RKHS

Reproducing Kernel Hilbert Spaces (RKHS) offer a very convenient framework for function approximation. Here

Definition: A Hilbert space of functions $D \longrightarrow \mathbb{C}$, $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$, is a RKHS if for all $\mathbf{x} \in D$, the evaluation functional $e_{\mathbf{x}} : f \in \mathcal{H} \longrightarrow f(\mathbf{x}) \in \mathbb{C}$ are continuous.

From the so-called *Riesz representation theorem*, for all $\mathbf{x} \in D$ there exists an element of $\mathcal{H}$, denoted here $k_{\mathbf{x}}$, such that $f(\mathbf{x}) = \langle f, k_{\mathbf{x}} \rangle_{\mathcal{H}}$.

From such a RKHS and the collection of Riesz evaluation representers $k_{\mathbf{x}}$, the "kernel" $k : D \times D \longrightarrow \mathbb{C}$ associated with $\mathcal{H}$ can be defined as follows:

$$k : (\mathbf{x}, \mathbf{x}') \in D \times D \longrightarrow k(\mathbf{x}, \mathbf{x}') = \langle k_{\mathbf{x}'}, k_{\mathbf{x}} \rangle_{\mathcal{H}}$$

Easy to check: k is a p.d. kernel.

RKHS

Reproducing Kernel Hilbert Spaces (RKHS) offer a very convenient framework for function approximation. Here

Definition: A Hilbert space of functions $D \longrightarrow \mathbb{C}$, $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$, is a RKHS if for all $\mathbf{x} \in D$, the evaluation functional $e_{\mathbf{x}} : f \in \mathcal{H} \longrightarrow f(\mathbf{x}) \in \mathbb{C}$ are continuous.

From the so-called *Riesz representation theorem*, for all $\mathbf{x} \in D$ there exists an element of $\mathcal{H}$, denoted here $k_{\mathbf{x}}$, such that $f(\mathbf{x}) = \langle f, k_{\mathbf{x}} \rangle_{\mathcal{H}}$.

From such a RKHS and the collection of Riesz evaluation representers $k_{\mathbf{x}}$, the "kernel" $k : D \times D \longrightarrow \mathbb{C}$ associated with $\mathcal{H}$ can be defined as follows:

$$k : (\mathbf{x}, \mathbf{x}') \in D \times D \longrightarrow k(\mathbf{x}, \mathbf{x}') = \langle k_{\mathbf{x}'}, k_{\mathbf{x}} \rangle_{\mathcal{H}}$$

Easy to check: k is a p.d. kernel.
Less easy to check: any p.d. kernel defines a unique RKHS
$\rightarrow$ Moore-Aronszajn theorem (Published 1950 :-)

Representing RKHSs based on the Mercer theorem

For simplicity, let us consider here a RKHS $\mathcal{H}_k$ associated with a real-valued Mercer kernel k. $\mathcal{H}_k$ can be represented more concretely as follows.

Representing RKHSs based on the Mercer theorem

For simplicity, let us consider here a RKHS $\mathcal{H}_k$ associated with a real-valued Mercer kernel k. $\mathcal{H}_k$ can be represented more concretely as follows.

$$\mathcal{H}_k = \left\{ f = \sum_{j=1}^{\infty} \alpha_j \sqrt{\lambda_j} \phi_j,\ \alpha \in \mathbb{R}^{\mathbb{N}} : \sum_{j=1}^{+\infty} \alpha_j^2 < \infty \right\}$$

with $\langle \sum_{j=1}^{\infty} \alpha_j \sqrt{\lambda_j} \phi_j, \sum_{j=1}^{\infty} \beta_j \sqrt{\lambda_j} \phi_j \rangle_{\mathcal{H}} := \sum_{j=1}^{\infty} \alpha_j \beta_j$.

Representing RKHSs based on the Mercer theorem

For simplicity, let us consider here a RKHS $\mathcal{H}_k$ associated with a real-valued Mercer kernel k. $\mathcal{H}_k$ can be represented more concretely as follows.

$$\mathcal{H}_k = \left\{ f = \sum_{j=1}^{\infty} \alpha_j \sqrt{\lambda_j} \phi_j,\ \alpha \in \mathbb{R}^{\mathbb{N}} : \sum_{j=1}^{+\infty} \alpha_j^2 < \infty \right\}$$

with $\langle \sum_{j=1}^{\infty} \alpha_j \sqrt{\lambda_j} \phi_j, \sum_{j=1}^{\infty} \beta_j \sqrt{\lambda_j} \phi_j \rangle_{\mathcal{H}} := \sum_{j=1}^{\infty} \alpha_j \beta_j$.

Comparing this with the K-L expansion of a GP with kernel k, we find that in the case of an infinite number of non-zero eigenvalues, the paths of Z are **not** in $\mathcal{H}_k$ with probability 1 (Parzen-Kallianpur-LePage theorem, as discussed in Lukić and Beder 2001). However, it can be shown that in general GP paths belong to bigger RKHSs (See, e.g., Steinwart 2017 for more detail).

Some properties of GRFs and kernels

Back to centred Z for simplicity, one can define a (pseudo-)metric d_Z on D by

$$d_Z^2(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(Z_{\mathbf{x}} - Z_{\mathbf{x}'})^2)\right] = k(\mathbf{x}, \mathbf{x}) + k(\mathbf{x}', \mathbf{x}') - 2k(\mathbf{x}, \mathbf{x}')$$

A number of properties of Z are driven by d_Z.

Some properties of GRFs and kernels

Back to centred Z for simplicity, one can define a (pseudo-)metric d_Z on D by

$$d_Z^2(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(Z_{\mathbf{x}} - Z_{\mathbf{x}'})^2)\right] = k(\mathbf{x}, \mathbf{x}) + k(\mathbf{x}', \mathbf{x}') - 2k(\mathbf{x}, \mathbf{x}')$$

A number of properties of Z are driven by d_Z. For instance,

Theorem (Sufficient condition for the continuity of GRF paths)

Let $(Z_{\mathbf{x}})_{\mathbf{x}\in D}$ be a separable Gaussian random field on a compact index set $D \subset \mathbb{R}^d$. If for some $0 < C < \infty$ and $\delta, \eta > 0$,

$$d_Z^2(\mathbf{x}, \mathbf{x}') \leq \frac{C}{\left|\log ||\mathbf{x} - \mathbf{x}'||\right|^{1+\delta}}$$

for all $\mathbf{x}, \mathbf{x}' \in D$ with $||\mathbf{x} - \mathbf{x}'|| < \eta$, then the paths of Z are almost surely continuous and bounded.

See, e.g., M. Scheuerer's PhD thesis (2009) for details.

Some properties of GRFs and kernels

Starting from p.d. kernels notably obtained via Bochner's theorem, an appealing approach to enrich them is by operations conserving symmetric positive definiteness.

Some properties of GRFs and kernels

Starting from p.d. kernels notably obtained via Bochner's theorem, an appealing approach to enrich them is by operations conserving symmetric positive definiteness.

Classical operations of that kind notably encompass:

- Non-negative linear combinations of p.d. kernels
- Products and tensor products of p.d. kernels
- Multiplication by $\sigma(\mathbf{x})\sigma(\mathbf{x}')$ for $\sigma : \mathbf{x} \in D \longrightarrow [0, +\infty)$
- Deformations/warpings: $k(g(\mathbf{x}), g(\mathbf{x}'))$ for $g : D \longrightarrow D$
- Convolutions, etc. . .

See, e.g., Section "making new kernels from old" of the book *Gaussian Processes for Machine Learning* (cited earlier).

The Branin-Hoo function

The rescaled Branin-Hoo function f is defined over $[0,1]^2$ by

$$f(x_1, x_2) = f_{\text{BH}}(15x_1 - 5, 15x_2),$$

where

$$f_{\text{BH}} : (x_1, x_2) \in [-5, 10] \times [0, 15] \longrightarrow a(x_2 - bx_1^2 + cx_1 - r) + s(1-t)\cos(x_1) + s,$$

with $a = 1, b = 5/(4\pi^2), c = 5/\pi, r = 6, s = 10$ and $t = 1/(8\pi)$ back.

Ordinary Kriging Equations –for completeness!–

Assume Z has a covariance kernel k, and constant mean $\mu \in \mathbb{R}$

Ordinary Kriging Equations –for completeness!–

Assume Z has a covariance kernel k, and constant mean $\mu \in \mathbb{R}$

$$\begin{cases} m_n(\mathbf{x}) & = k(\mathbf{X}_n, \mathbf{x})^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} \mathbf{z}_n + \widehat{\mu}_n (1 - k(\mathbf{X}_n, \mathbf{x})^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} \mathbb{1}_n) \\ k_n(\mathbf{x}, \mathbf{x}') & = k(\mathbf{x}, \mathbf{x}') - k(\mathbf{X}_n, \mathbf{x})^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x}') \\ & \quad + \frac{\left(1 - \mathbb{1}_n^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x})\right)\left(1 - \mathbb{1}_n^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x}')\right)}{\left(\mathbb{1}_n^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} \mathbb{1}_n\right)} \end{cases}$$

Ordinary Kriging Equations –for completeness!–

Assume Z has a covariance kernel k, and constant mean $\mu \in \mathbb{R}$

$$\begin{cases} m_n(\mathbf{x}) & = k(\mathbf{X}_n, \mathbf{x})^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} \mathbf{z}_n + \widehat{\mu}_n (1 - k(\mathbf{X}_n, \mathbf{x})^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} \mathbb{1}_n) \\ k_n(\mathbf{x}, \mathbf{x}') & = k(\mathbf{x}, \mathbf{x}') - k(\mathbf{X}_n, \mathbf{x})^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x}') \\ & \quad + \frac{\left(1 - \mathbb{1}_n^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x})\right)\left(1 - \mathbb{1}_n^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x}')\right)}{\left(\mathbb{1}_n^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} \mathbb{1}_n\right)} \end{cases}$$

where $\widehat{\mu}_n = \frac{\mathbb{1}_n^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} \mathbf{z}_n}{(\mathbb{1}_n^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} \mathbb{1}_n)}$.

Ordinary Kriging Equations –for completeness!–

Assume Z has a covariance kernel k, and constant mean $\mu \in \mathbb{R}$

$$\begin{cases} m_n(\mathbf{x}) & = k(\mathbf{X}_n, \mathbf{x})^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} \mathbf{z}_n + \widehat{\mu}_n (1 - k(\mathbf{X}_n, \mathbf{x})^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} \mathbb{1}_n) \\ k_n(\mathbf{x}, \mathbf{x}') & = k(\mathbf{x}, \mathbf{x}') - k(\mathbf{X}_n, \mathbf{x})^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x}') \\ & \quad + \frac{\left(1 - \mathbb{1}_n^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x})\right)\left(1 - \mathbb{1}_n^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x}')\right)}{\left(\mathbb{1}_n^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} \mathbb{1}_n\right)} \end{cases}$$

where $\widehat{\mu}_n = \frac{\mathbb{1}_n^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} \mathbf{z}_n}{(\mathbb{1}_n^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} \mathbb{1}_n)}$.

Under standard conditions, m_n **and** k_n **are** Z**'s conditional mean and covariance** and

$$\boxed{\mathcal{L}(Z | Z_{X_n} = \mathbf{z}_n) = \mathcal{GRF}\left(m_n(\cdot), k_n(\cdot, \cdot')\right)}$$

back

Heterogeneously noisy OK Equations

$$\begin{cases} m_n(\mathbf{x}) & = \widehat{\mu}_n + \mathbf{k}_n(\mathbf{x})^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} \left(\widetilde{\mathbf{z}}_n - \widehat{\mu}_n \mathbb{1}_n\right) \\ k_n(\mathbf{x}, \mathbf{x}') & = k(\mathbf{x}, \mathbf{x}') - k(\mathbf{X}_n, \mathbf{x})^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} k(\mathbf{X}_n, \mathbf{x}') \\ & \quad + \frac{\left(1 - \mathbb{1}_n^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} k(\mathbf{X}_n, \mathbf{x})\right)\left(1 - \mathbb{1}_n^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} k(\mathbf{X}_n, \mathbf{x}')\right)}{\left(\mathbb{1}_n^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} \mathbb{1}_n\right)} \end{cases}$$

Heterogeneously noisy OK Equations

$$\begin{cases} m_n(\mathbf{x}) & = \widehat{\mu}_n + \mathbf{k}_n(\mathbf{x})^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} \left(\widetilde{\mathbf{z}}_n - \widehat{\mu}_n \mathbb{1}_n\right) \\ k_n(\mathbf{x}, \mathbf{x}') & = k(\mathbf{x}, \mathbf{x}') - k(\mathbf{X}_n, \mathbf{x})^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} k(\mathbf{X}_n, \mathbf{x}') \\ & \quad + \frac{\left(1 - \mathbb{1}_n^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} k(\mathbf{X}_n, \mathbf{x})\right)\left(1 - \mathbb{1}_n^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} k(\mathbf{X}_n, \mathbf{x}')\right)}{\left(\mathbb{1}_n^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} \mathbb{1}_n\right)} \end{cases}$$

where $\widehat{\mu}_n = \frac{\mathbb{1}_n^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} \widetilde{\mathbf{z}}_n}{(\mathbb{1}_n^T (\mathbf{K} + \Delta_n)^{-1} \mathbb{1}_n)}$.

Heterogeneously noisy OK Equations

$$\begin{cases} m_n(\mathbf{x}) & = \widehat{\mu}_n + \mathbf{k}_n(\mathbf{x})^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} \left(\widetilde{\mathbf{z}}_n - \widehat{\mu}_n \mathbb{1}_n\right) \\ k_n(\mathbf{x}, \mathbf{x}') & = k(\mathbf{x}, \mathbf{x}') - k(\mathbf{X}_n, \mathbf{x})^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} k(\mathbf{X}_n, \mathbf{x}') \\ & \quad + \frac{\left(1 - \mathbb{1}_n^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} k(\mathbf{X}_n, \mathbf{x})\right)\left(1 - \mathbb{1}_n^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} k(\mathbf{X}_n, \mathbf{x}')\right)}{\left(\mathbb{1}_n^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} \mathbb{1}_n\right)} \end{cases}$$

where $\widehat{\mu}_n = \frac{\mathbb{1}_n^T (k(\mathbf{X}_n, \mathbf{X}_n) + \Delta_n)^{-1} \widetilde{\mathbf{z}}_n}{(\mathbb{1}_n^T (\mathbf{K} + \Delta_n)^{-1} \mathbb{1}_n)}$.

Under usual assumptions, and if Z and the $\varepsilon_i' s$ are independent:

$$\boxed{\mathcal{L}(Z | \widetilde{A_n}) = \mathcal{N}\left(m_n(\cdot), k_n(\cdot, \cdot')\right)}$$

back