Gaussian processes for regression, global optimization and set estimation

David Ginsbourger [1,2]

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

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

Gaussian processes and Uncertainty Quantification Summer School

Sheffield, 12-15 September 2016

Part I

Introduction

Set up

Goal: estimate a deterministic function $f : \mathbf{x} \in E \mapsto f(\mathbf{x}) \in F$ and/or quantities relying on it based on a limited number of evaluations of f.

Set up

Goal: estimate a deterministic function $f : \mathbf{x} \in E \mapsto f(\mathbf{x}) \in F$ and/or quantities relying on it based on a limited number of evaluations of f.

Often, $\mathbf{x}$ is living in a compact subspace D of $E = \mathbb{R}^d$ ($d \geq 1$) and the response space is $F = \mathbb{R}^k$ ($k \geq 1$). Here $k = 1$.

Set up

Goal: estimate a deterministic function $f : \mathbf{x} \in E \mapsto f(\mathbf{x}) \in F$ and/or quantities relying on it based on a limited number of evaluations of f.

Often, $\mathbf{x}$ is living in a compact subspace D of $E = \mathbb{R}^d$ ($d \geq 1$) and the response space is $F = \mathbb{R}^k$ ($k \geq 1$). Here $k = 1$.

Two typical examples where f stems from numerical simulations

- Safety engineering: $\mathbf{x}$ is a vector parametrizing some system and f returns an indicator of dangerousness. It is then crucial to understand which $\mathbf{x}$'s lead to "high" values of $f(\mathbf{x})$.

Set up

Goal: estimate a deterministic function $f : \mathbf{x} \in E \mapsto f(\mathbf{x}) \in F$ and/or quantities relying on it based on a limited number of evaluations of f.

Often, $\mathbf{x}$ is living in a compact subspace D of $E = \mathbb{R}^d$ $(d \geq 1)$ and the response space is $F = \mathbb{R}^k$ $(k \geq 1)$. Here $k = 1$.

Two typical examples where f stems from numerical simulations

- Safety engineering: $\mathbf{x}$ is a vector parametrizing some system and f returns an indicator of dangerousness. It is then crucial to understand which $\mathbf{x}$'s lead to "high" values of $f(\mathbf{x})$.
- Flow simulation: $\mathbf{x}$ stands e.g. for the medium, boundary conditions, etc. and f returns the evolution of a fluid and/or a measure of discrepancy between simulation results and given observation results.

Set up

Typical situation : f was evaluated at a set of "points" $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D \subset E$ and one wishes to estimate a quantity relying on f and/or run new evaluations in order to improve this estimation.

Set up

Typical situation : f was evaluated at a set of "points" $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D \subset E$ and one wishes to estimate a quantity relying on f and/or run new evaluations in order to improve this estimation.

$\Rightarrow$ legitimate to rely on some approximation(s) of f knowing $f(\mathbf{x}_i) + \epsilon_i$ $(1 \leq i \leq n)$. A number of approaches do exist...

Set up

Typical situation : f was evaluated at a set of "points" $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D \subset E$ and one wishes to estimate a quantity relying on f and/or run new evaluations in order to improve this estimation.

$\Rightarrow$ legitimate to rely on some approximation(s) of f knowing $f(\mathbf{x}_i) + \epsilon_i$ $(1 \leq i \leq n)$. A number of approaches do exist...

Principles of the Gaussian Process approach (GP): suppose that, *a priori*, f is a realization of a GP $(Z_{\mathbf{x}})_{\mathbf{x} \in D}$ and approximate f and/or the quantities of interest via the **conditional distribution** of Z knowing $Z_{\mathbf{x}_i} + \varepsilon_i = f(\mathbf{x}_i) + \epsilon_i$.

Set up

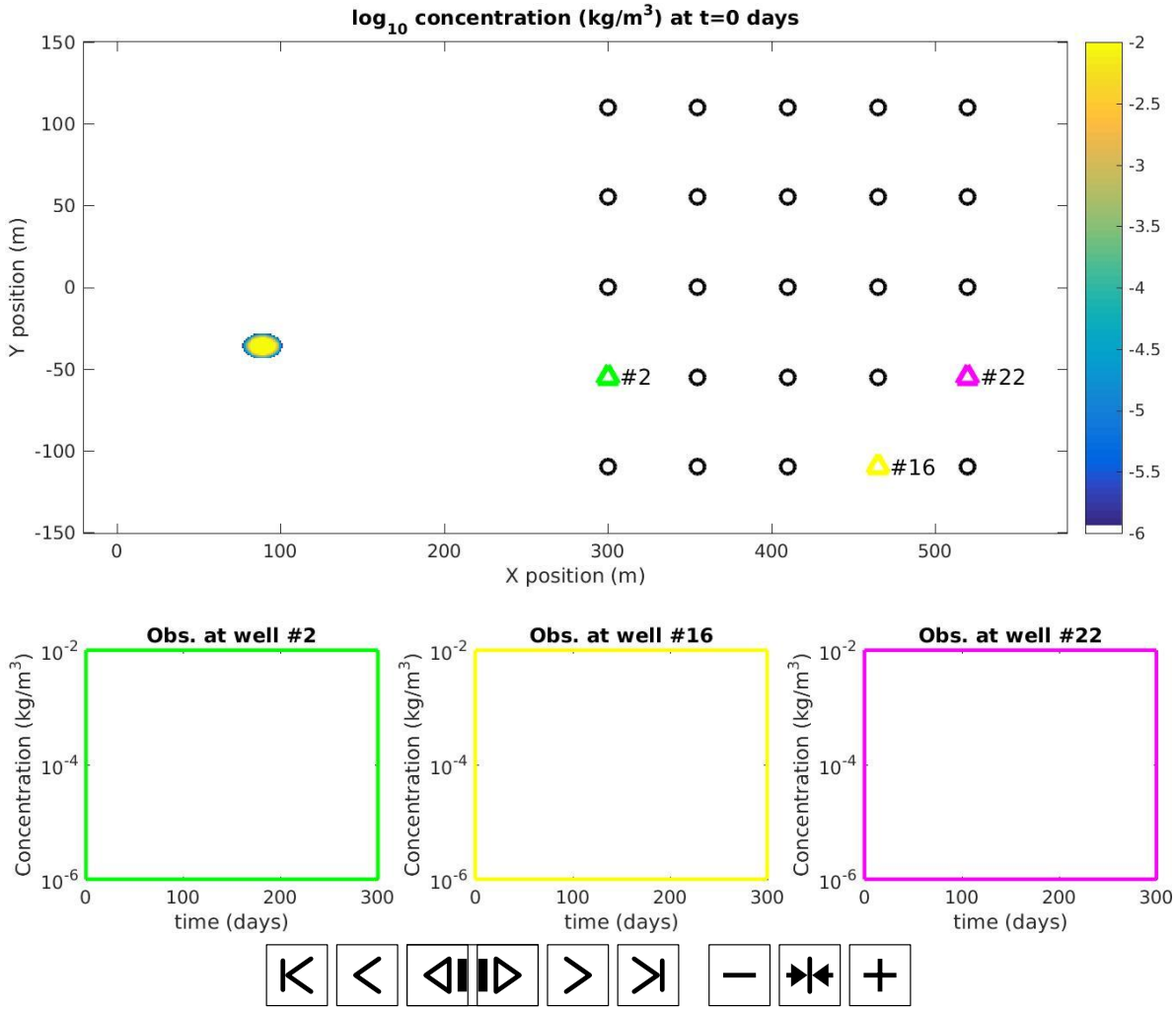
Typical situation : f was evaluated at a set of "points" $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D \subset E$ and one wishes to estimate a quantity relying on f and/or run new evaluations in order to improve this estimation.

$\Rightarrow$ legitimate to rely on some approximation(s) of f knowing $f(\mathbf{x}_i) + \epsilon_i$ $(1 \leq i \leq n)$. A number of approaches do exist...

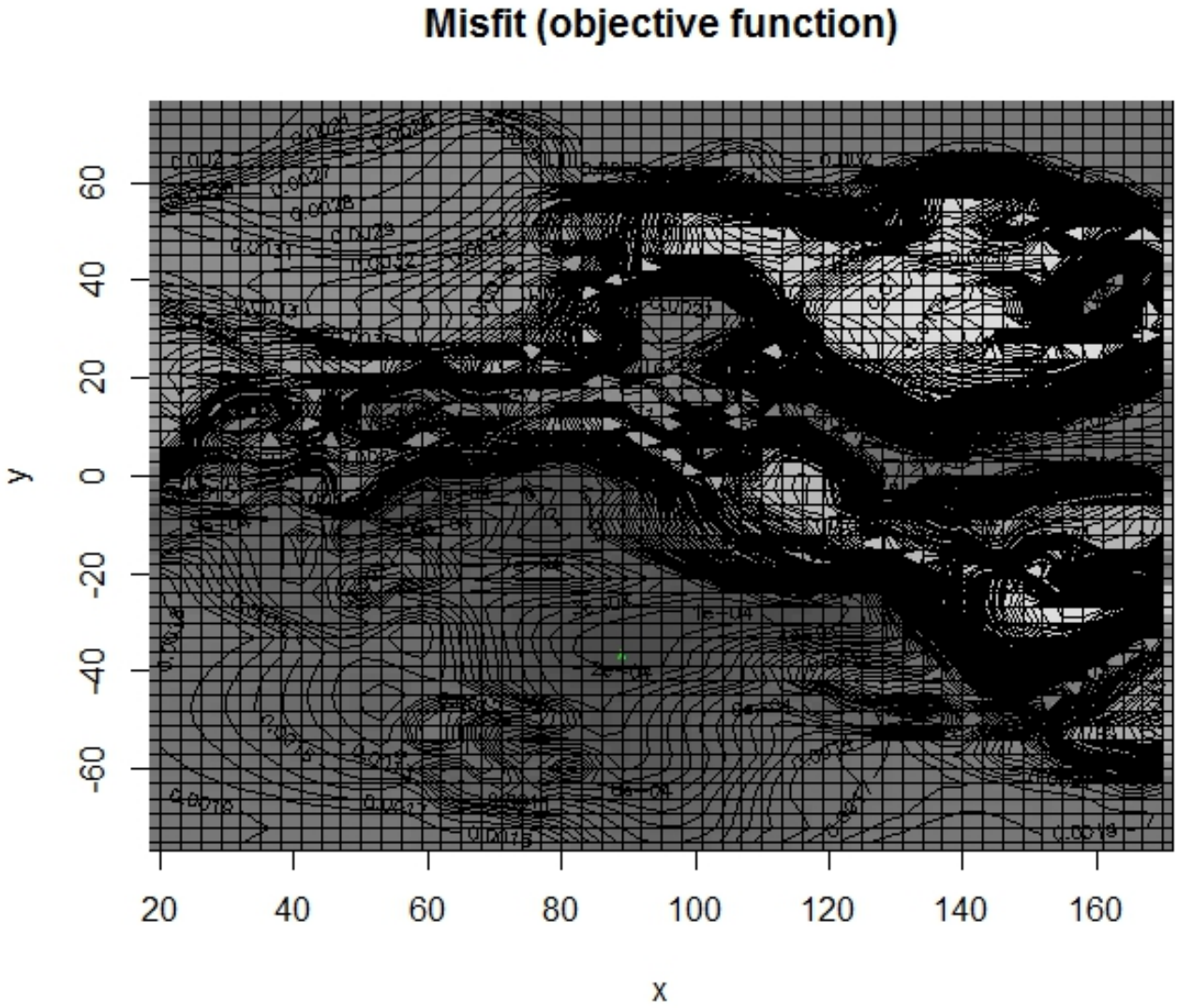
Principles of the Gaussian Process approach (GP): suppose that, *a priori*, f is a realization of a GP $(Z_{\mathbf{x}})_{\mathbf{x} \in D}$ and approximate f and/or the quantities of interest via the **conditional distribution** of Z knowing $Z_{\mathbf{x}_i} + \varepsilon_i = f(\mathbf{x}_i) + \epsilon_i$.

$\Rightarrow$ very practical for sequential design of experiments.

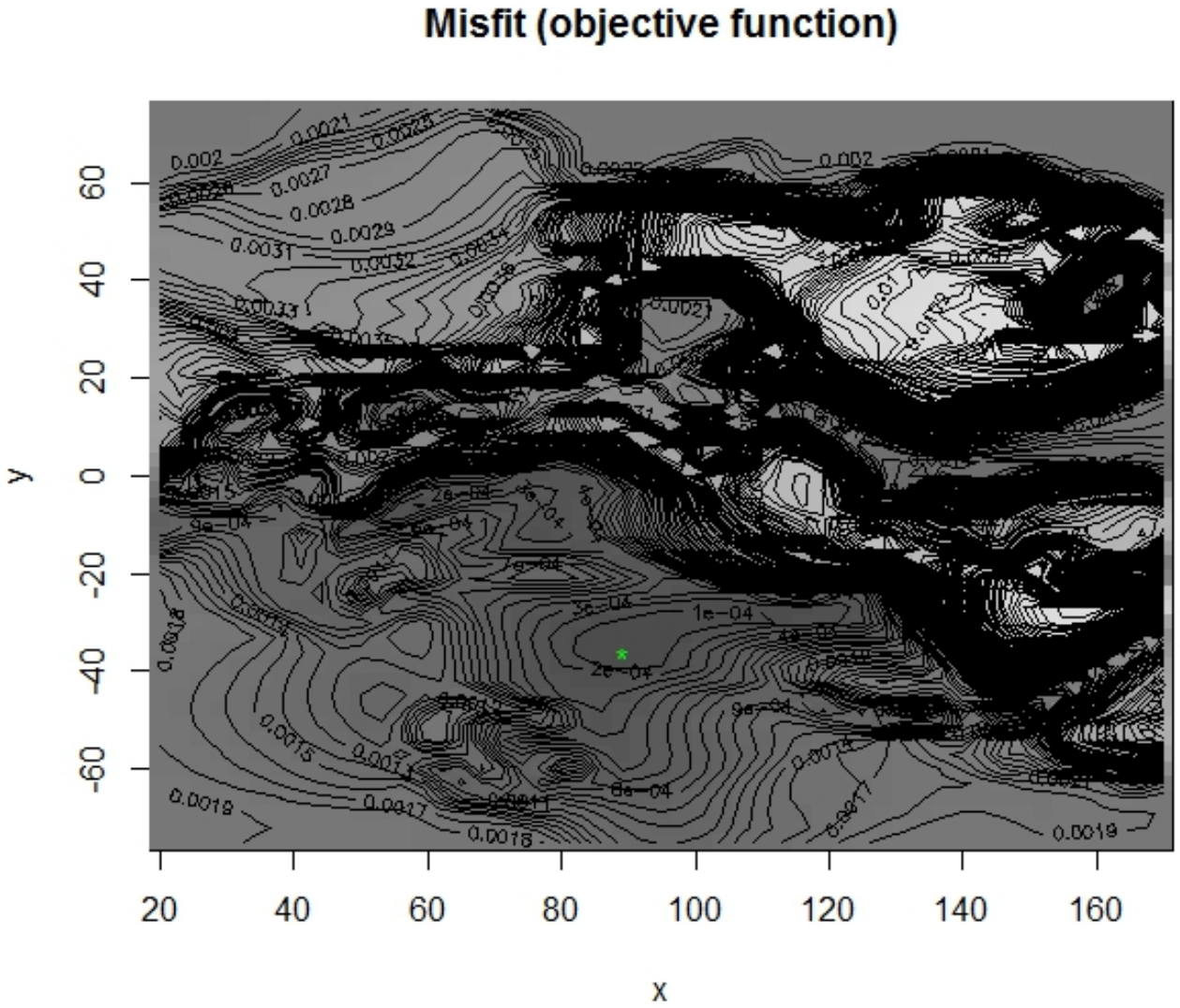
Example: inverse problem in hydrogeology



A costly full factorial experimental design!

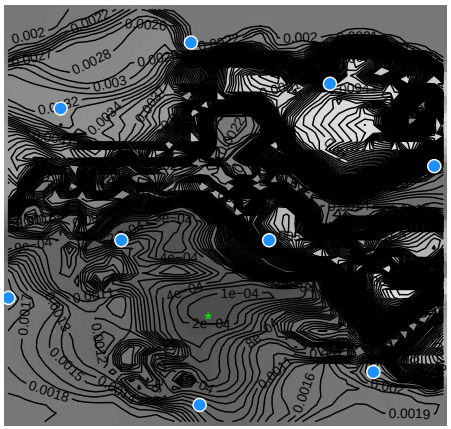


A costly full factorial experimental design!

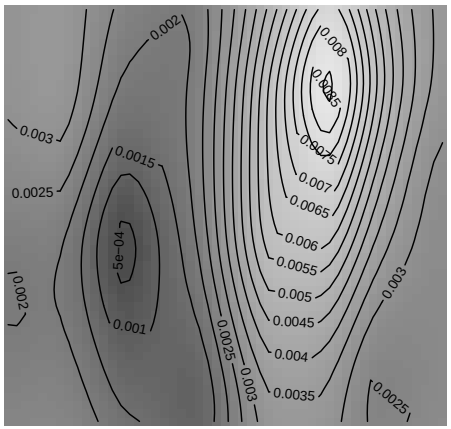


Source localization by Bayesian optimization

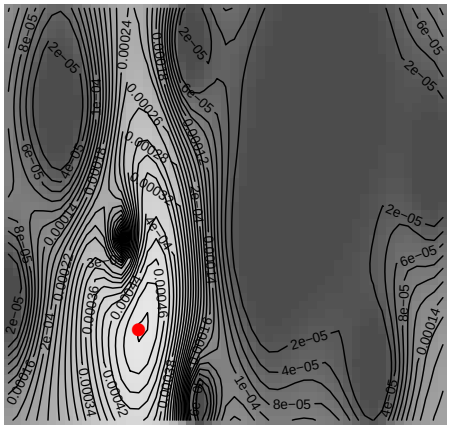
Misfit (objective function)



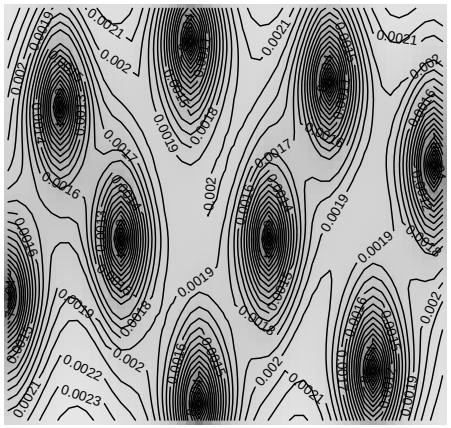
GP mean prediction



Expected Improvement



GP standard deviation



The previous example was produced in the framework of an ongoing collaboration with T. Krityakierne (now at Mahidol University, Bangkok), G. Pirot (University of Lausanne), and P. Renard (University of Neuchâtel).

The previous example was produced in the framework of an ongoing collaboration with T. Krityakierne (now at Mahidol University, Bangkok), G. Pirot (University of Lausanne), and P. Renard (University of Neuchâtel).

A few general questions

- About global optimization: does it converge? Does it parallelize well? Can it be applied in higher dimensions? In noise?

The previous example was produced in the framework of an ongoing collaboration with T. Krityakierne (now at Mahidol University, Bangkok), G. Pirot (University of Lausanne), and P. Renard (University of Neuchâtel).

A few general questions

- About global optimization: does it converge? Does it parallelize well? Can it be applied in higher dimensions? In noise? $\Rightarrow$ [Part II]

The previous example was produced in the framework of an ongoing collaboration with T. Krityakierne (now at Mahidol University, Bangkok), G. Pirot (University of Lausanne), and P. Renard (University of Neuchâtel).

A few general questions

- About global optimization: does it converge? Does it parallelize well? Can it be applied in higher dimensions? In noise? $\Rightarrow$ [Part II]
- What if the target is not to recover (nearly) optimal points but other quantities such as excursion sets or their measure?

The previous example was produced in the framework of an ongoing collaboration with T. Krityakierne (now at Mahidol University, Bangkok), G. Pirot (University of Lausanne), and P. Renard (University of Neuchâtel).

A few general questions

- About global optimization: does it converge? Does it parallelize well? Can it be applied in higher dimensions? In noise? $\Rightarrow$ [Part II]
- What if the target is not to recover (nearly) optimal points but other quantities such as excursion sets or their measure? $\Rightarrow$ [Part III]

The previous example was produced in the framework of an ongoing collaboration with T. Krityakierne (now at Mahidol University, Bangkok), G. Pirot (University of Lausanne), and P. Renard (University of Neuchâtel).

A few general questions

- About global optimization: does it converge? Does it parallelize well? Can it be applied in higher dimensions? In noise? $\Rightarrow$ [Part II]
- What if the target is not to recover (nearly) optimal points but other quantities such as excursion sets or their measure? $\Rightarrow$ [Part III]
- What kind of mathematical properties of f can be incorporated and or learnt with Gaussian Process approaches?

The previous example was produced in the framework of an ongoing collaboration with T. Krityakierne (now at Mahidol University, Bangkok), G. Pirot (University of Lausanne), and P. Renard (University of Neuchâtel).

A few general questions

- About global optimization: does it converge? Does it parallelize well? Can it be applied in higher dimensions? In noise? $\Rightarrow$ [Part II]
- What if the target is not to recover (nearly) optimal points but other quantities such as excursion sets or their measure? $\Rightarrow$ [Part III]
- What kind of mathematical properties of f can be incorporated and or learnt with Gaussian Process approaches? $\Rightarrow$ [Part IV]

The previous example was produced in the framework of an ongoing collaboration with T. Krityakierne (now at Mahidol University, Bangkok), G. Pirot (University of Lausanne), and P. Renard (University of Neuchâtel).

A few general questions

- About global optimization: does it converge? Does it parallelize well? Can it be applied in higher dimensions? In noise? $\Rightarrow$ [Part II]
- What if the target is not to recover (nearly) optimal points but other quantities such as excursion sets or their measure? $\Rightarrow$ [Part III]
- What kind of mathematical properties of f can be incorporated and or learnt with Gaussian Process approaches? $\Rightarrow$ [Part IV]

Let us start by a short reminder about Gaussian Processes.

Preliminary: priors on functions?

A real-valued random field Z with index set D is a collection of random variables $(Z_{\mathbf{x}})_{\mathbf{x}\in D}$ defined over the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

Preliminary: priors on functions?

A real-valued random field 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 random field are defined through their finite-dimensional distributions, that is joint distributions of random vectors of the form $(Z_{\mathbf{x}_1}, \ldots, Z_{\mathbf{x}_n})$ for any finite set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset D$ $(n \geq 1)$.

Preliminary: priors on functions?

A real-valued random field 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 random field are defined through their finite-dimensional distributions, that is joint distributions of random vectors of the form $(Z_{\mathbf{x}_1}, \ldots, Z_{\mathbf{x}_n})$ for any finite set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset D$ $(n \geq 1)$.

Kolmogorov's extension theorem tells us that families of joint probability distributions satisfying a few consistency conditions define random fields.

Preliminary: priors on functions?

A real-valued random field 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 random field are defined through their finite-dimensional distributions, that is joint distributions of random vectors of the form $(Z_{\mathbf{x}_1}, \ldots, Z_{\mathbf{x}_n})$ for any finite set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset D$ $(n \geq 1)$.

Kolmogorov's extension theorem tells us that families of joint probability distributions satisfying a few consistency conditions define random fields.

Gaussian Random Fields (GRFs, a.k.a. GPs here)

One major example of such family is given by *multivariate Gaussian distributions*.

Preliminary: priors on functions?

A real-valued random field 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 random field are defined through their finite-dimensional distributions, that is joint distributions of random vectors of the form $(Z_{\mathbf{x}_1}, \ldots, Z_{\mathbf{x}_n})$ for any finite set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset D$ $(n \geq 1)$.

Kolmogorov's extension theorem tells us that families of joint probability distributions satisfying a few consistency conditions define random fields.

Gaussian Random Fields (GRFs, a.k.a. GPs here)

One major example of such family is given by *multivariate Gaussian distributions*. By specifying mean and covariance matrix of random vectors corresponding to any finite set of locations, one defines a **GRF**/**GP**.

Preliminary: GPs

Hence a GP Z is completely defined -as random element over the cylindrical σ algebra of $\mathbb{R}^D$- 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 its law is characterized by

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

Preliminary: GPs

Hence a GP Z is completely defined -as random element over the cylindrical σ algebra of $\mathbb{R}^D$- 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 its law is characterized by

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

While m 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 a covariance matrix, i.e. symmetric positive semi-definite, for any set of points. k satisfying such property are refereed to as **p.d. kernels**.

Preliminary: GPs

Hence a GP Z is completely defined -as random element over the cylindrical σ algebra of $\mathbb{R}^D$- 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 its law is characterized by

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

While m 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 a covariance matrix, i.e. symmetric positive semi-definite, for any set of points. k satisfying such property are refereed 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”).

Preliminary: Examples of p.d. kernels and GRFs

For $d = 1$ and $k(t, t') = \min(t, t')$, one gets the Brownian Motion $(W_t)_{t\in[0,1]}$.

Still for $d = 1$, $k(t, t') = \min(t, t') \times (1 - \max(t, t'))$ gives the so-called Brownian Bridge, say $(B_t)_{t\in[0,1]}$.

Also, for $H \in (0, 1)$, $k(t, t') = \frac{1}{2}(|t|^{2H} + |t'|^{2H} - |t - t'|^{2H})$ is the covariance kernel of the *fractional (or "fractal") Brownian Motion* with Hurst coefficient H.

Preliminary: Examples of p.d. kernels and GRFs

For $d = 1$ and $k(t, t') = \min(t, t')$, one gets the Brownian Motion $(W_t)_{t\in[0,1]}$.

Still for $d = 1$, $k(t, t') = \min(t, t') \times (1 - \max(t, t'))$ gives the so-called Brownian Bridge, say $(B_t)_{t\in[0,1]}$.

Also, for $H \in (0, 1)$, $k(t, t') = \frac{1}{2}(|t|^{2H} + |t'|^{2H} - |t - t'|^{2H})$ is the covariance kernel of the *fractional (or "fractal") Brownian Motion* with Hurst coefficient H.

$k(t, t') = e^{-|t-t'|}$ is called exponential kernel and characterizes the *Ornstein-Uhlenbeck process*. $k(t, t') = e^{-|t-t'|^2}$ is the Gaussian kernel.

Preliminary: Examples of p.d. kernels and GRFs

For $d = 1$ and $k(t, t') = \min(t, t')$, one gets the Brownian Motion $(W_t)_{t\in[0,1]}$.

Still for $d = 1$, $k(t, t') = \min(t, t') \times (1 - \max(t, t'))$ gives the so-called Brownian Bridge, say $(B_t)_{t\in[0,1]}$.

Also, for $H \in (0, 1)$, $k(t, t') = \frac{1}{2}(|t|^{2H} + |t'|^{2H} - |t - t'|^{2H})$ is the covariance kernel of the *fractional (or "fractal") Brownian Motion* with Hurst coefficient H.

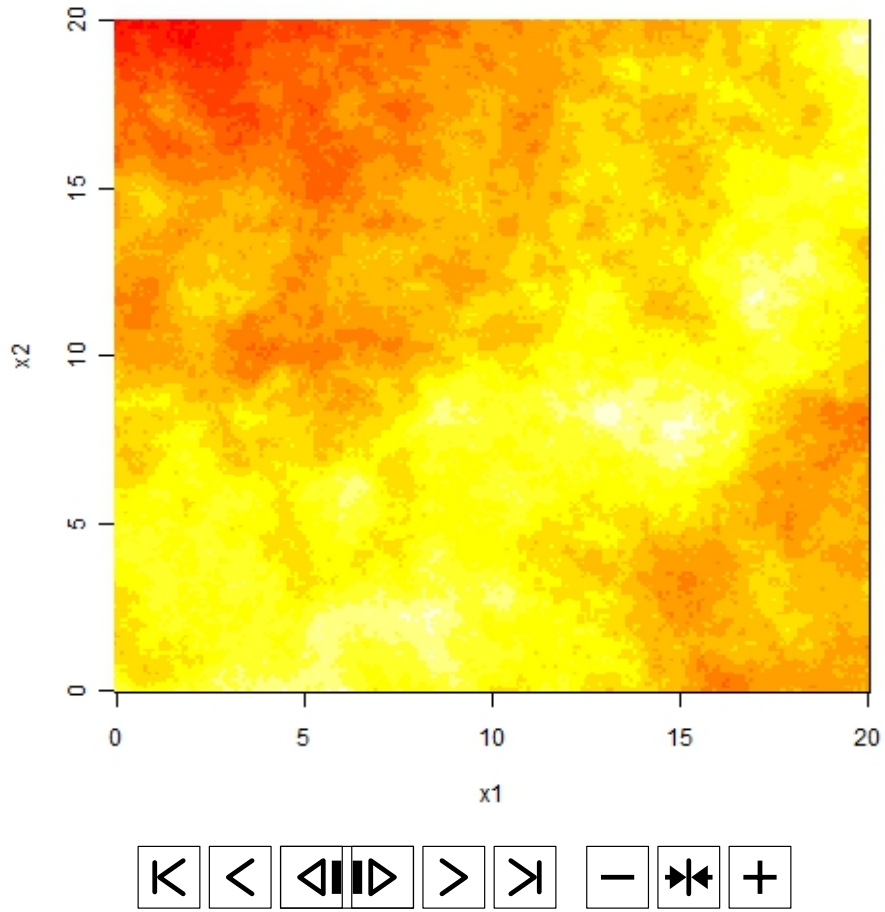
$k(t, t') = e^{-|t-t'|}$ is called exponential kernel and characterizes the *Ornstein-Uhlenbeck process*. $k(t, t') = e^{-|t-t'|^2}$ is the Gaussian kernel.

The two last kernels possess a so-called *stationarity* (or "shift-invariance") property. Also, in turns out that these kernels can be generalized to $d \geq 1$:

$$k(\mathbf{x}, \mathbf{x}') = e^{-||\mathbf{x}-\mathbf{x}'||} \quad \text{("isotropic exponential")}$$

$$k(\mathbf{x}, \mathbf{x}') = e^{-||\mathbf{x}-\mathbf{x}'||^2} \quad \text{("isotropic Gaussian")}$$

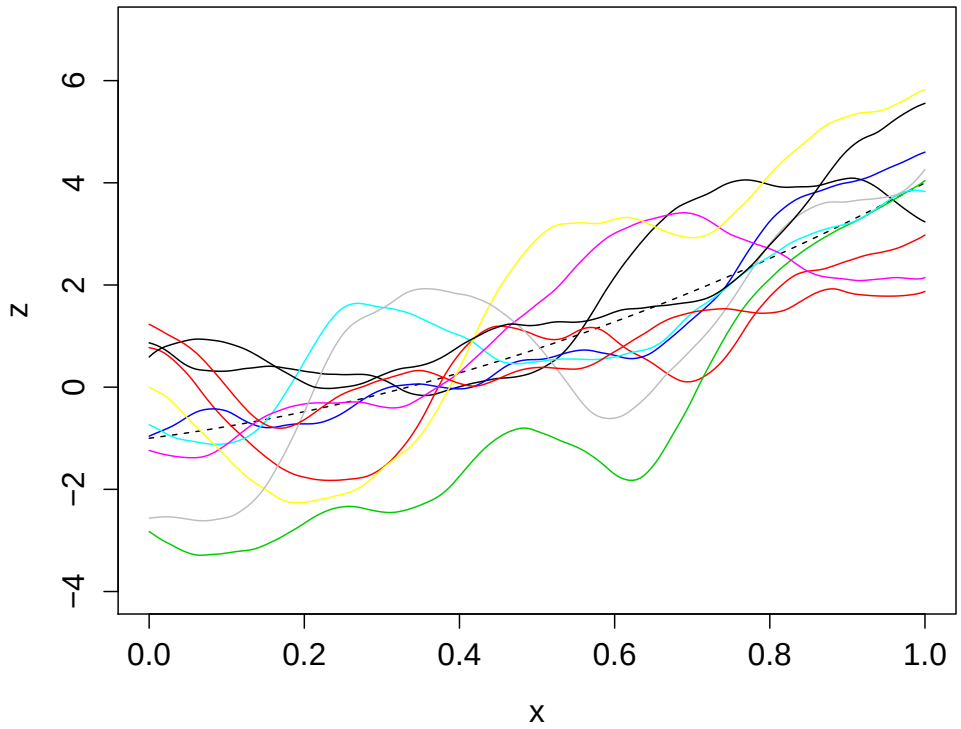
Some GRF R simulations (d=2) with *RandomFields*



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

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

Furthermore, here trend is a trend $m(t) = -1 + 2t + 3t^2$.



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.

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.

Properties of GRFs and kernels

Several other pathwise properties of Z can be controlled through k, such as differentiability, but also symmetries, harmonicity, and more (Cf. Part IV).

Properties of GRFs and kernels

Several other pathwise properties of Z can be controlled through k, such as differentiability, but also symmetries, harmonicity, and more (Cf. Part IV).

In practice, the choice of k often relies (in)directly on Bochner's theorem. Noting $k(\mathbf{h}) = k(\mathbf{x}, \mathbf{x}')$ for k stationary and $\mathbf{h} = \mathbf{x} - \mathbf{x}'$, we have

Properties of GRFs and kernels

Several other pathwise properties of Z can be controlled through k, such as differentiability, but also symmetries, harmonicity, and more (Cf. Part IV).

In practice, the choice of k often relies (in)directly on Bochner's theorem. Noting $k(\mathbf{h}) = k(\mathbf{x}, \mathbf{x}')$ for k stationary and $\mathbf{h} = \mathbf{x} - \mathbf{x}'$, we have

Theorem (Bochner's theorem)

A function $k : \mathbb{R}^d \longrightarrow \mathbb{R}$ is continuous and positive definite if and only if a finite symmetric non-negative measure ν on $\mathbb{R}^d$ exists so that

$$k(\mathbf{h}) = \int_{\mathbb{R}^d} \cos(\langle \mathbf{h}, \mathbf{w} \rangle) \nu(\mathrm{d}\mathbf{w}) \quad \textit{for all } \mathbf{h} \in \mathbb{R}^d$$

ν is then called the spectral measure of k.

Properties of GRFs and kernels

Several other pathwise properties of Z can be controlled through k, such as differentiability, but also symmetries, harmonicity, and more (Cf. Part IV).

In practice, the choice of k often relies (in)directly on Bochner's theorem. Noting $k(\mathbf{h}) = k(\mathbf{x}, \mathbf{x}')$ for k stationary and $\mathbf{h} = \mathbf{x} - \mathbf{x}'$, we have

Theorem (Bochner's theorem)

A function $k : \mathbb{R}^d \longrightarrow \mathbb{R}$ is continuous and positive definite if and only if a finite symmetric non-negative measure ν on $\mathbb{R}^d$ exists so that

$$k(\mathbf{h}) = \int_{\mathbb{R}^d} \cos(\langle \mathbf{h}, \mathbf{w} \rangle) \nu(\mathrm{d}\mathbf{w}) \quad \textit{for all } \mathbf{h} \in \mathbb{R}^d$$

ν is then called the spectral measure of k.

For ν absolutely continuous with $\nu = \varphi \mathrm{dLeb}_d$, φ is called the spectral density of k. Example: Matérn $\equiv \varphi(\mathbf{w}) = (1 + ||\mathbf{w}||^2)^{-r}$.

Properties of GRFs and kernels

Sarting from known d.p. kernels, it is common to enrich the choice by appealing to operations preserving symmetry & positive definiteness, e.g.:

- 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. . .

Properties of GRFs and kernels

Sarting from known d.p. kernels, it is common to enrich the choice by appealing to operations preserving symmetry & positive definiteness, e.g.:

- 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. . .

C. E. Rasmussen and C.K.I. Williams (2006).
Gaussian Processes for Machine Learning.
Section “making new kernels from old”.
MIT Press

A few 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.

A few 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.

Part II

About GP-based Bayesian optimization

Some seminal papers

H.J. Kushner (1964).

A new method of locating the maximum of an arbitrary multi-peak curve in the presence of noise.

Journal of Basic Engineering, 86:97-106.

J. Mockus (1972).

On Bayesian methods for seeking the extremum.

Automatics and Computers (Avtomatika i Vychislitel'naya Tekhnika), 4(1):53-62.

J. Mockus, V. Tiesis, and A. Zilinskas (1978).

The application of Bayesian methods for seeking the extremum.

In Dixon, L. C. W. and Szegö, G. P., editors, Towards Global Optimisation, volume 2, pages 117-129. Elsevier Science Ltd., North Holland, Amsterdam.

J.M. Calvin (1997).

Average performance of a class of adaptive algorithms for global optimization.

The Annals of Applied Probability, 7(3):711-730.

M. Schonlau, W.J. Welch and D.R. Jones (1998).

Efficient Global Optimization of Expensive Black-box Functions.

Journal of Global Optimization.

Decision-theoretic roots of EI (1)

Assume that f (modelled by Z) was already evaluated at a set of points $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset D$ ($n \geq n_0$), at that one wishes to perform additional evaluations at one or more points $\mathbf{x}_{n+j} \in D$ ($1 \leq j \leq q$, $q \geq 1$).

Decision-theoretic roots of EI (1)

Assume that f (modelled by Z) was already evaluated at a set of points $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset D$ ($n \geq n_0$), at that one wishes to perform additional evaluations at one or more points $\mathbf{x}_{n+j} \in D$ ($1 \leq j \leq q$, $q \geq 1$).

A rather natural score to judge performances in minimization at step n is the **regret** $t_n - f^\star$ (a.k.a. optimality gap), where $t_n = \min_{1 \leq i \leq n} f(\mathbf{x}_i)$.

Decision-theoretic roots of EI (1)

Assume that f (modelled by Z) was already evaluated at a set of points $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset D$ ($n \geq n_0$), at that one wishes to perform additional evaluations at one or more points $\mathbf{x}_{n+j} \in D$ ($1 \leq j \leq q$, $q \geq 1$).

A rather natural score to judge performances in minimization at step n is the **regret** $t_n - f^\star$ (a.k.a. optimality gap), where $t_n = \min_{1 \leq i \leq n} f(\mathbf{x}_i)$.

When choosing $\mathbf{x}_{n+j} \in D$ ($1 \leq j \leq q$), we wish them to minimize $t_{n+q} - f^\star$.

Decision-theoretic roots of EI (1)

Assume that f (modelled by Z) was already evaluated at a set of points $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset D$ ($n \geq n_0$), at that one wishes to perform additional evaluations at one or more points $\mathbf{x}_{n+j} \in D$ ($1 \leq j \leq q$, $q \geq 1$).

A rather natural score to judge performances in minimization at step n is the **regret** $t_n - f^\star$ (a.k.a. optimality gap), where $t_n = \min_{1 \leq i \leq n} f(\mathbf{x}_i)$.

When choosing $\mathbf{x}_{n+j} \in D$ ($1 \leq j \leq q$), we wish them to minimize $t_{n+q} - f^\star$.
Two problems arise:

1. t_{n+q} cannot be known before evaluating f at the new points
2. $f^\star$ is generally not known at all

Decision-theoretic roots of EI (2)

Capitalizing quantities where f is replaced by Z, the standard approach to deal with the first problem is to minimize the expected (simple) regret

$$(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+q}) \in D^q \longrightarrow \mathbb{E}_n \left[T_{n+q} - Z^\star \right],$$

where $\mathbb{E}_n$ refers to the expectation conditional on $\{Z(\mathbf{X}_n) = \mathbf{z}_n\}$.

Decision-theoretic roots of EI (2)

Capitalizing quantities where f is replaced by Z, the standard approach to deal with the first problem is to minimize the expected (simple) regret

$$(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+q}) \in D^q \longrightarrow \mathbb{E}_n \left[T_{n+q} - Z^\star \right],$$

where $\mathbb{E}_n$ refers to the expectation conditional on $\{Z(\mathbf{X}_n) = \mathbf{z}_n\}$.

That $Z^\star$ is unknown can be circumvented since minimizing $\mathbb{E}_n \left[T_{n+q} - Z^\star \right]$ is equivalent to minimizing $\mathbb{E}_n \left[T_{n+q} \right]$ or $\mathbb{E}_n \left[T_{n+q} - T_n \right]$.

Decision-theoretic roots of EI (2)

Capitalizing quantities where f is replaced by Z, the standard approach to deal with the first problem is to minimize the expected (simple) regret

$$(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+q}) \in D^q \longrightarrow \mathbb{E}_n \left[T_{n+q} - Z^\star \right],$$

where $\mathbb{E}_n$ refers to the expectation conditional on $\{Z(\mathbf{X}_n) = \mathbf{z}_n\}$.

That $Z^\star$ is unknown can be circumvented since minimizing $\mathbb{E}_n \left[T_{n+q} - Z^\star \right]$ is equivalent to minimizing $\mathbb{E}_n \left[T_{n+q} \right]$ or $\mathbb{E}_n \left[T_{n+q} - T_n \right]$. Besides this,

$$T_n - T_{n+q} = \left(T_n - \min_{1 \leq j \leq q} Z_{\mathbf{x}_{n+j}} \right)^+ .$$

Decision-theoretic roots of EI (2)

Capitalizing quantities where f is replaced by Z, the standard approach to deal with the first problem is to minimize the expected (simple) regret

$$(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+q}) \in D^q \longrightarrow \mathbb{E}_n \left[T_{n+q} - Z^\star \right],$$

where $\mathbb{E}_n$ refers to the expectation conditional on $\{Z(\mathbf{X}_n) = \mathbf{z}_n\}$.

That $Z^\star$ is unknown can be circumvented since minimizing $\mathbb{E}_n \left[T_{n+q} - Z^\star \right]$ is equivalent to minimizing $\mathbb{E}_n \left[T_{n+q} \right]$ or $\mathbb{E}_n \left[T_{n+q} - T_n \right]$. Besides this,

$$T_n - T_{n+q} = \left(T_n - \min_{1 \leq j \leq q} Z_{\mathbf{x}_{n+j}} \right)^+ .$$

Hence, **minimizing the expected regret is equivalent to maximizing**

$$\mathbb{E}_n \left[\left(T_n - \min_{1 \leq j \leq q} Z_{\mathbf{x}_{n+j}} \right)^+ \right].$$

Definition and derivation of EI

Setting $q = 1$, the **Expected Improvement** criterion at step n is defined as:

$$\mathrm{EI}_n : \mathbf{x} \in D \longrightarrow \mathrm{EI}_n(\mathbf{x}) = \mathbb{E}_n[(T_n - Z_{\mathbf{x}})^+].$$

Definition and derivation of EI

Setting $q = 1$, the **Expected Improvement** criterion at step n is defined as:

$$\mathrm{EI}_n : \mathbf{x} \in D \longrightarrow \mathrm{EI}_n(\mathbf{x}) = \mathbb{E}_n[(T_n - Z_{\mathbf{x}})^+].$$

As $T_n = t_n$ and $Z_{\mathbf{x}} \sim \mathcal{N}(m_n(\mathbf{x}), k_n(\mathbf{x}, \mathbf{x}))$ conditionally on $\{Z(\mathbf{X}_n) = \mathbf{z}_n\}$,

$$\mathrm{EI}_n(\mathbf{x}) = \begin{cases} 0 & \text{if } s_n(\mathbf{x}) = 0 \\ s_n(\mathbf{x}) \{u_n(\mathbf{x})\Phi(u_n(\mathbf{x})) + \phi(u_n(\mathbf{x}))\} & \text{else.} \end{cases}$$

where $s_n(\mathbf{x}) = \sqrt{k_n(\mathbf{x}, \mathbf{x})}$ and $u_n(\mathbf{x}) = (t_n - m_n(\mathbf{x}))/s_n(\mathbf{x})$.

Definition and derivation of EI

Setting $q = 1$, the **Expected Improvement** criterion at step n is defined as:

$$\mathrm{EI}_n : \mathbf{x} \in D \longrightarrow \mathrm{EI}_n(\mathbf{x}) = \mathbb{E}_n[(T_n - Z_{\mathbf{x}})^+].$$

As $T_n = t_n$ and $Z_{\mathbf{x}} \sim \mathcal{N}(m_n(\mathbf{x}), k_n(\mathbf{x}, \mathbf{x}))$ conditionally on $\{Z(\mathbf{X}_n) = \mathbf{z}_n\}$,

$$\mathrm{EI}_n(\mathbf{x}) = \begin{cases} 0 & \text{if } s_n(\mathbf{x}) = 0 \\ s_n(\mathbf{x}) \{u_n(\mathbf{x})\Phi(u_n(\mathbf{x})) + \phi(u_n(\mathbf{x}))\} & \text{else.} \end{cases}$$

where $s_n(\mathbf{x}) = \sqrt{k_n(\mathbf{x}, \mathbf{x})}$ and $u_n(\mathbf{x}) = (t_n - m_n(\mathbf{x}))/s_n(\mathbf{x})$.

N.B.: EI_n is a first order moment of a truncated Gaussian.

Selected properties of the EI criterion

1. EI_n is non-negative over D and vanishes at $\boldsymbol{X}_n$
2. EI_n is generally not convex/concave and highly multi-modal
3. The regularity of EI_n is driven by k_n
4. If k possesses the "No-Empty-Ball" property more, sampling using EI eventually fills the space provided f belongs to the RKHS of kernel k.

Selected properties of the EI criterion

1. EI_n is non-negative over D and vanishes at $\boldsymbol{X}_n$
2. EI_n is generally not convex/concave and highly multi-modal
3. The regularity of EI_n is driven by k_n
4. If k possesses the "No-Empty-Ball" property more, sampling using EI eventually fills the space provided f belongs to the RKHS of kernel k.

NB: new convergence results for EI and more are presented in

J. Bect, F. Bachoc and D. Ginsbourger (2016).
A supermartingale approach to Gaussian process based sequential design of experiments.
HAL/Arxiv paper (hal-01351088, Arxiv: 1608.01118).

Parallelizing EI algorithms with the multipoint EI

Extending the standard EI to $q > 1$ points is of practical interest as it allows distributing EI algorithms over several processors/computers in parallel.

Parallelizing EI algorithms with the multipoint EI

Extending the standard EI to $q > 1$ points is of practical interest as it allows distributing EI algorithms over several processors/computers in parallel.

Efforts have recently been paid to calculate the "multipoint EI" criterion:

$$\mathrm{EI}_n : (\mathbf{x}_1, \ldots, \mathbf{x}_q) \in D^q \longrightarrow \mathbb{E}_n \left[\left(T_n - \min_{1 \leq j \leq q} (Z_{\mathbf{x}_j}) \right)^+ \right].$$

Parallelizing EI algorithms with the multipoint EI

Extending the standard EI to $q > 1$ points is of practical interest as it allows distributing EI algorithms over several processors/computers in parallel.

Efforts have recently been paid to calculate the "multipoint EI" criterion:

$$\mathrm{EI}_n : (\mathbf{x}_1, \ldots, \mathbf{x}_q) \in D^q \longrightarrow \mathbb{E}_n \left[\left(T_n - \min_{1 \leq j \leq q} (Z_{\mathbf{x}_j}) \right)^+ \right].$$

D. Ginsbourger, R. Le Riche, L. Carraro (2010)

Kriging is well-suited to parallelize optimization

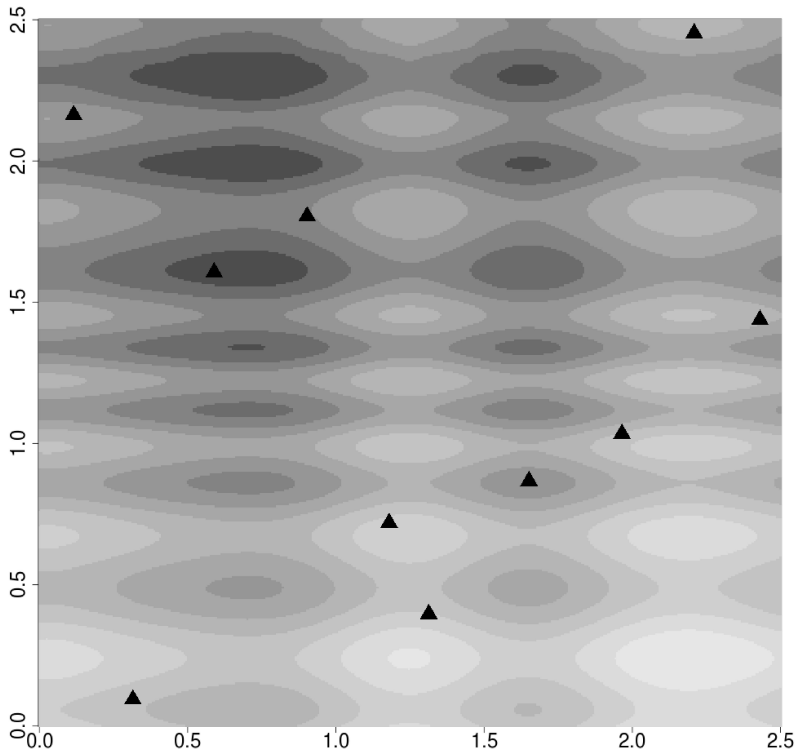
In Computational Intelligence in Expensive Optimization Problems, Adaptation Learning and Optimization, pages 131-162. Springer Berlin Heidelberg, 2010

C. Chevalier, D. Ginsbourger (2013) Goto equations

Fast computation of the multipoint Expected Improvement with applications in batch selection.

Learning and Intelligent Optimization (LION7)

Multipoint EI: Example



Multipoint EI: Latest results and ongoing work

The main computational bottleneck lies in the maximization of the Multipoint EI criterion over D^q.

Multipoint EI: Latest results and ongoing work

The main computational bottleneck lies in the maximization of the Multipoint EI criterion over D^q.

Closed formulae as well as fast and efficient approximations have been obtained for the gradient of the multipoint EI criterion.

S. Marmin, C. Chevalier, D. Ginsbourger. (2015).
Differentiating the multipoint Expected Improvement for optimal batch design.
International Workshop on Machine learning, Optimization and big Data.

Multipoint EI: Latest results and ongoing work

The main computational bottleneck lies in the maximization of the Multipoint EI criterion over D^q.

Closed formulae as well as fast and efficient approximations have been obtained for the gradient of the multipoint EI criterion.

S. Marmin, C. Chevalier, D. Ginsbourger. (2015).
Differentiating the multipoint Expected Improvement for optimal batch design.
International Workshop on Machine learning, Optimization and big Data.

S. Marmin, C. Chevalier, D. Ginsbourger. (2016+).
Efficient batch-sequential Bayesian optimization with moments of truncated Gaussian vectors.
Hal/Arxiv paper (https://hal.archives-ouvertes.fr/hal-01361894/).

Multipoint EI: Latest results and ongoing work

The main computational bottleneck lies in the maximization of the Multipoint EI criterion over D^q.

Closed formulae as well as fast and efficient approximations have been obtained for the gradient of the multipoint EI criterion.

S. Marmin, C. Chevalier, D. Ginsbourger. (2015).
Differentiating the multipoint Expected Improvement for optimal batch design.
International Workshop on Machine learning, Optimization and big Data.

S. Marmin, C. Chevalier, D. Ginsbourger. (2016+).
Efficient batch-sequential Bayesian optimization with moments of truncated Gaussian vectors.
Hal/Arxiv paper (https://hal.archives-ouvertes.fr/hal-01361894/).

N.B.: alternative approaches for maximizing the multipoint EI, that rely on stochastic gradients, have been developed by Peter Frazier and his group.

On finite-time Bayesian Global Optimization

Let us now assume that a fixed number of evaluations (after step n_0), say $r \geq 1$, is allocated for the sequential minimization of f (one point at a time).

By construction, we know that EI is optimal at the last iteration. However, maximizing EI is generally not optimal if there remains more than one step.

On finite-time Bayesian Global Optimization

Let us now assume that a fixed number of evaluations (after step n_0), say $r \geq 1$, is allocated for the sequential minimization of f (one point at a time).

By construction, we know that EI is optimal at the last iteration. However, maximizing EI is generally not optimal if there remains more than one step.

There exists in fact an optimal strategy, relying on backward induction. Taking a simple example with $r = 2$, the optimal action at step n_0 is to maximize

$$\mathbf{x} \longrightarrow \mathbb{E}_{n_0}\left[\left(T_{n_0} - \min(Z_{\mathbf{x}}, Z_{X_2^*})\right)^+\right],$$

where X_2^* maximizes EI_{n_0+1} (and so depends on $Z_{\mathbf{x}}$).

On finite-time BO: a few references

J. Mockus (1982)

The Bayesian approach to global optimization.

In Systems Modeling and Optimization, volume 38, pp. 473-481. Springer.

M. A. Osborne, R. Garnett, and S.J. Roberts (2009)

Gaussian processes for global optimization.

Learning and Intelligent OptimizatioN conference (LION3).

D. Ginsbourger, R. Le Riche (2010)

Towards Gaussian process-based optimization with finite time horizon.

mODa 9 Advances in Model-Oriented Design and Analysis, Contributions to Statistics, pages 89-96. Physica-Verlag HD.

S. Grünewälder, J. Y. Audibert, M. Opper, and J. Shawe-Taylor (2010).

Regret bounds for Gaussian process bandit problems. Goto detail.

In International Conference on Artificial Intelligence and Statistics (pp. 273-280), MIT Press.

J. Gonzalez, M. Osborne, N. Lawrence (2016).

GLASSES: Relieving The Myopia Of Bayesian Optimisation.

In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (pp. 790-799).

About (very) high-dimensional BO

One of the bottlenecks of Global Optimization is high-dimensionality. How to minimize f when n is severely limited and d is very large?

One often (realistically) assume that f only depends on $d_e << d$ variables.

About (very) high-dimensional BO

One of the bottlenecks of Global Optimization is high-dimensionality. How to minimize f when n is severely limited and d is very large?

One often (realistically) assume that f only depends on $d_e << d$ variables.

Some attempts have recently been done in Bayesian Optimization, that mostly rely on one of the two following ideas:

- Trying to identify the subset of d_e influential variables
- Restricting the search to one or more d_e-dimensional space(s) via random embedding(s)

About (very) high-dimensional BO: a few references

B. Chen, R.M. Castro, and A. Krause (2012)

Joint Optimization and Variable Selection of High-dimensional Gaussian Processes

In International Conference on Machine Learning

Z. Wang, M. Zoghi, F. Hutter, D. Matheson, and N. de Freitas (2013)

Bayesian optimization in a billion dimensions via random embeddings.

In International Joint Conferences on Artificial Intelligence

J. Djolonga, A. Krause, and V. Cevher (2013).

High-Dimensional Gaussian Process Bandits.

In Neural Information Processing Systems.

M. Binois, D. Ginsbourger, O. Roustant (2015).

A warped kernel improving robustness in Bayesian optimization via random embeddings.

In Learning and Intelligent Optimization (LION9)

M. Binois (2015).

Uncertainty quantification on Pareto fronts and high-dimensional strategies in Bayesian optimization, with applications in multi-objective automotive design.

Ph.D. thesis, Ecole des Mines de Saint-Etienne.

Mitigating model uncertainty in BO

Model-based criteria such as EI are usually calculated under the assumption that k and/or its parameters is/are known.

Mitigating model uncertainty in BO

Model-based criteria such as EI are usually calculated under the assumption that k and/or its parameters is/are known.

Incorporating estimation uncertainty into Bayesian global optimization algorithms has been done using various approaches, including notably

- Making it ”Full Bayesian”
- Appealing to parametric bootstrap

Mitigating model uncertainty in BO

Model-based criteria such as EI are usually calculated under the assumption that k and/or its parameters is/are known.

Incorporating estimation uncertainty into Bayesian global optimization algorithms has been done using various approaches, including notably

- Making it ”Full Bayesian”
- Appealing to parametric bootstrap

Calculating EI in this way was reported to favour exploratory behaviours.

Mitigating model uncertainty in BO: a few references


D. Ginsbourger, C. Helbert, L. Carraro (2008).
Discrete mixtures of kernels for Kriging-based Optimization.
Quality and Reliability Engineering International, 24(6):681-691.


R.B. Gramacy , M. Taddy (2010).
Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an R package for treed Gaussian process models.
Journal of Statistical Software, 33(6).


J.P.C. Kleijnen, W. van Beers, I. van Nieuwenhuyse (2012).
Expected improvement in efficient global optimization through bootstrapped kriging.
Journal of Global Optimization, 54(1):59-73.


R. Benassi, J. Bect, and E. Vazquez (2012).
Bayesian optimization using sequential Monte Carlo.
Learning and Intelligent Optimization (LION6).

Transition

A few other topics beyond our scope

- Multi-objective/constrained/robust Bayesian optimization $\Rightarrow$ Cf. notably works of Emmerich et al., Picheny et al., Binois et al., Féliot et al., etc.

Transition

A few other topics beyond our scope

- Multi-objective/constrained/robust Bayesian optimization $\Rightarrow$ Cf. notably works of Emmerich et al., Picheny et al., Binois et al., Féliot et al., etc.
- Multi-fidelity Bayesian optimization $\Rightarrow$ Cf. notably works of Forrester et al., Le Gratiet et al., Perdikaris et al., etc.

Transition

A few other topics beyond our scope

- Multi-objective/constrained/robust Bayesian optimization $\Rightarrow$ Cf. notably works of Emmerich et al., Picheny et al., Binois et al., Féliot et al., etc.
- Multi-fidelity Bayesian optimization $\Rightarrow$ Cf. notably works of Forrester et al., Le Gratiet et al., Perdikaris et al., etc.

A few references about the noisy aspect can be found here.

Transition

A few other topics beyond our scope

- Multi-objective/constrained/robust Bayesian optimization $\Rightarrow$ Cf. notably works of Emmerich et al., Picheny et al., Binois et al., Féliot et al., etc.
- Multi-fidelity Bayesian optimization $\Rightarrow$ Cf. notably works of Forrester et al., Le Gratiet et al., Perdikaris et al., etc.

A few references about the noisy aspect can be found here.

Note also that beyond EI, further criteria are in use for Bayesian optimization and related. Upper Confidence Bound strategies are quite popular, notably as they lead to elegant theoretical results; See e.g. works by Andreas Krause and team, and also recent contributions by Emile Contal et al.

Part III

On the estimation of excursion sets and their measure, and stepwise uncertainty reduction

Background and motivations

A number of practical problems boil down to determining sets of the form

$$\Gamma^\star = \{\mathbf{x} \in D : f(\mathbf{x}) \in T\} = f^{-1}(T)$$

where D is a compact subset of $\mathbb{R}^d$ ($d \geq 1$), $f : D \longrightarrow \mathbb{R}^k$ ($k \geq 1$) is a $(\mathcal{B}(D), \mathcal{B}(\mathbb{R}^k))$-measurable function, and $T \in \mathcal{B}(\mathbb{R}^k)$.

Background and motivations

A number of practical problems boil down to determining sets of the form

$$\Gamma^\star = \{\mathbf{x} \in D : f(\mathbf{x}) \in T\} = f^{-1}(T)$$

where D is a compact subset of $\mathbb{R}^d$ ($d \geq 1$), $f : D \longrightarrow \mathbb{R}^k$ ($k \geq 1$) is a $(\mathcal{B}(D), \mathcal{B}(\mathbb{R}^k))$-measurable function, and $T \in \mathcal{B}(\mathbb{R}^k)$.

For simplicity, we essentially focus today on the case where $k = 1$, f is continuous, and $T = [t, +\infty)$ for some prescribed $t \in \mathbb{R}$.

$\Gamma^\star = \{\mathbf{x} \in D : f(\mathbf{x}) \geq t\}$ is then referred to as the excursion set of f above t.

Background and motivations

A number of practical problems boil down to determining sets of the form

$$\Gamma^\star = \{\mathbf{x} \in D : f(\mathbf{x}) \in T\} = f^{-1}(T)$$

where D is a compact subset of $\mathbb{R}^d$ ($d \geq 1$), $f : D \longrightarrow \mathbb{R}^k$ ($k \geq 1$) is a $(\mathcal{B}(D), \mathcal{B}(\mathbb{R}^k))$-measurable function, and $T \in \mathcal{B}(\mathbb{R}^k)$.

For simplicity, we essentially focus today on the case where $k = 1$, f is continuous, and $T = [t, +\infty)$ for some prescribed $t \in \mathbb{R}$.

$\Gamma^\star = \{\mathbf{x} \in D : f(\mathbf{x}) \geq t\}$ is then referred to as the excursion set of f above t.

Our aim is to estimate $\Gamma^\star$ and quantify uncertainty on it when f can solely be evaluated at a few points $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset D$.

Test case from safety engineering

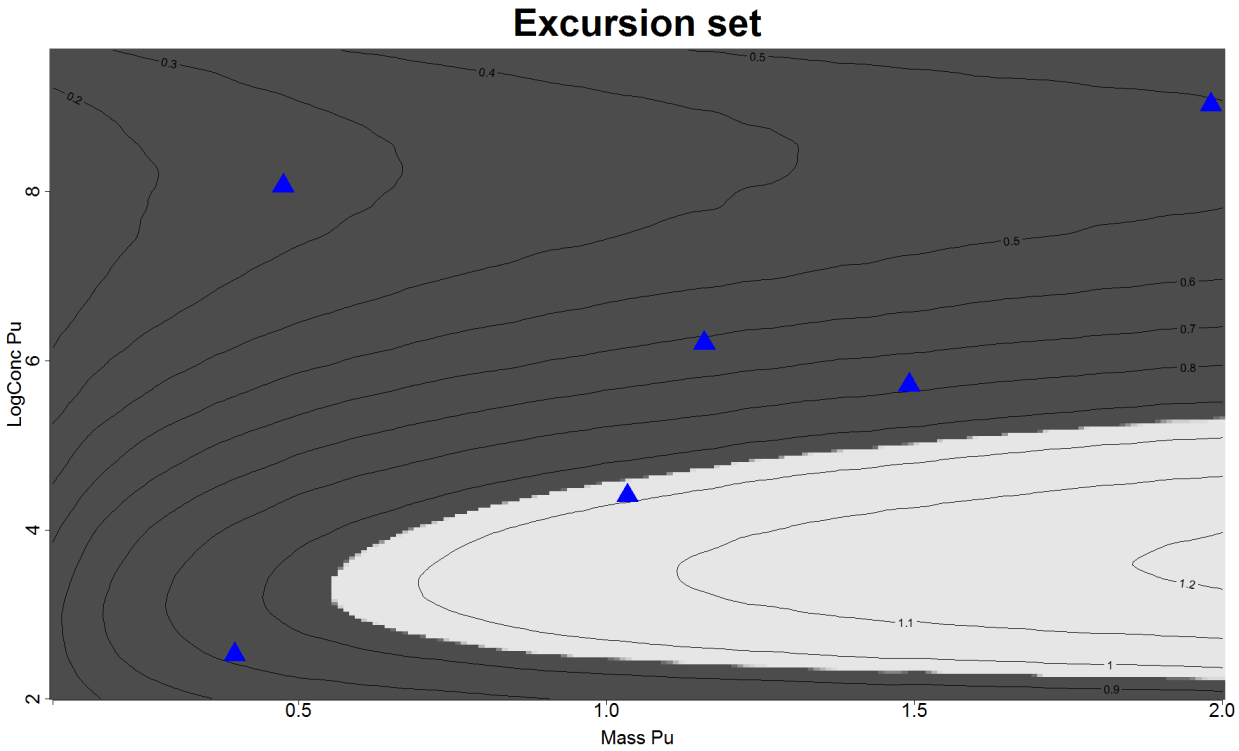


Figure: Excursion set (light gray) of a nuclear criticality safety coefficient depending on two design parameters. Blue triangles: initial experiments.

C. Chevalier (2013).
Fast uncertainty reduction strategies relying on Gaussian process models.
Ph.D. thesis, University of Bern.

Making a sensible estimation of $\Gamma^\star$ based on a drastically limited number of evaluations $f(\mathbf{X}_n) = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))'$ calls for additional assumptions on f.

Making a sensible estimation of $\Gamma^\star$ based on a drastically limited number of evaluations $f(\mathbf{X}_n) = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))'$ calls for additional assumptions on f.

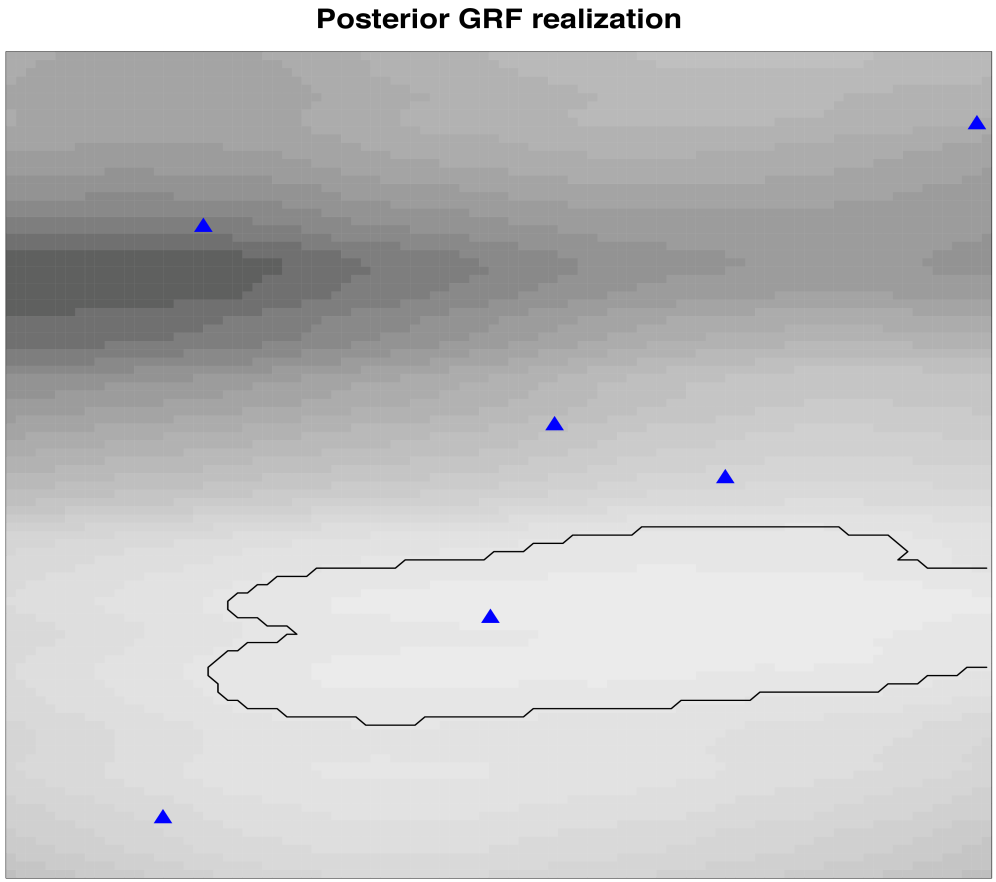
In the GP set-up, the main object of interest is represented by

$$\Gamma = \{\mathbf{x} \in D : Z(\mathbf{x}) \in T\} = Z^{-1}(T)$$

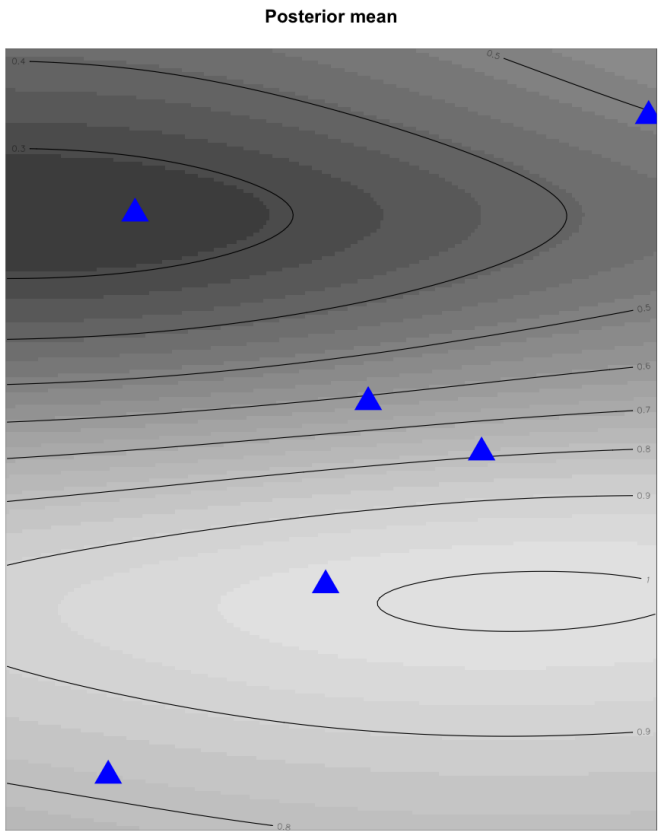
Under our previous assumptions on T (and assuming that is chosen Z with continuous paths a.s.), Γ appears to be a Random Closed Set.

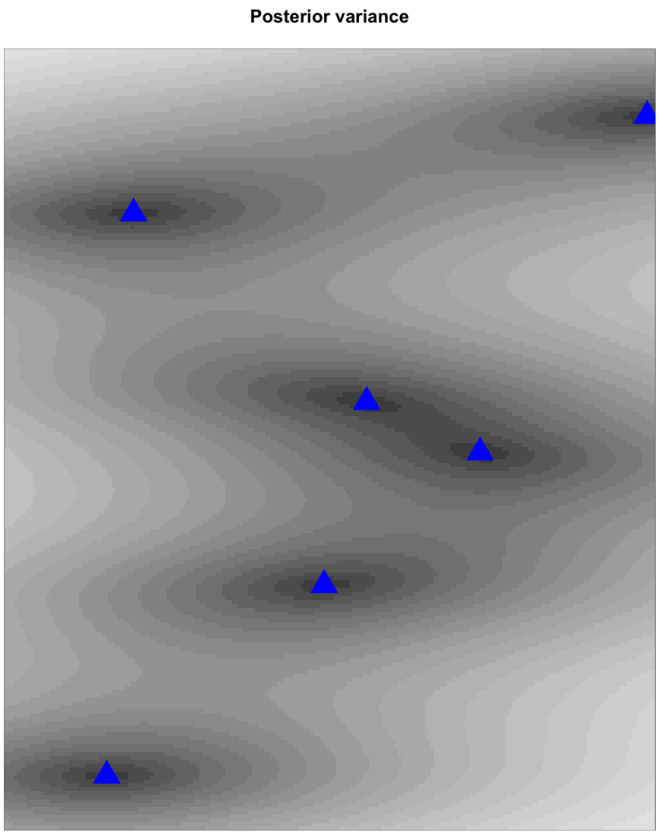
Simulating excursion sets under a GRF model

Several realizations of Γ simulated on a grid 50×50 knowing $Z(\mathbf{X}_n) = f(\mathbf{X}_n)$.

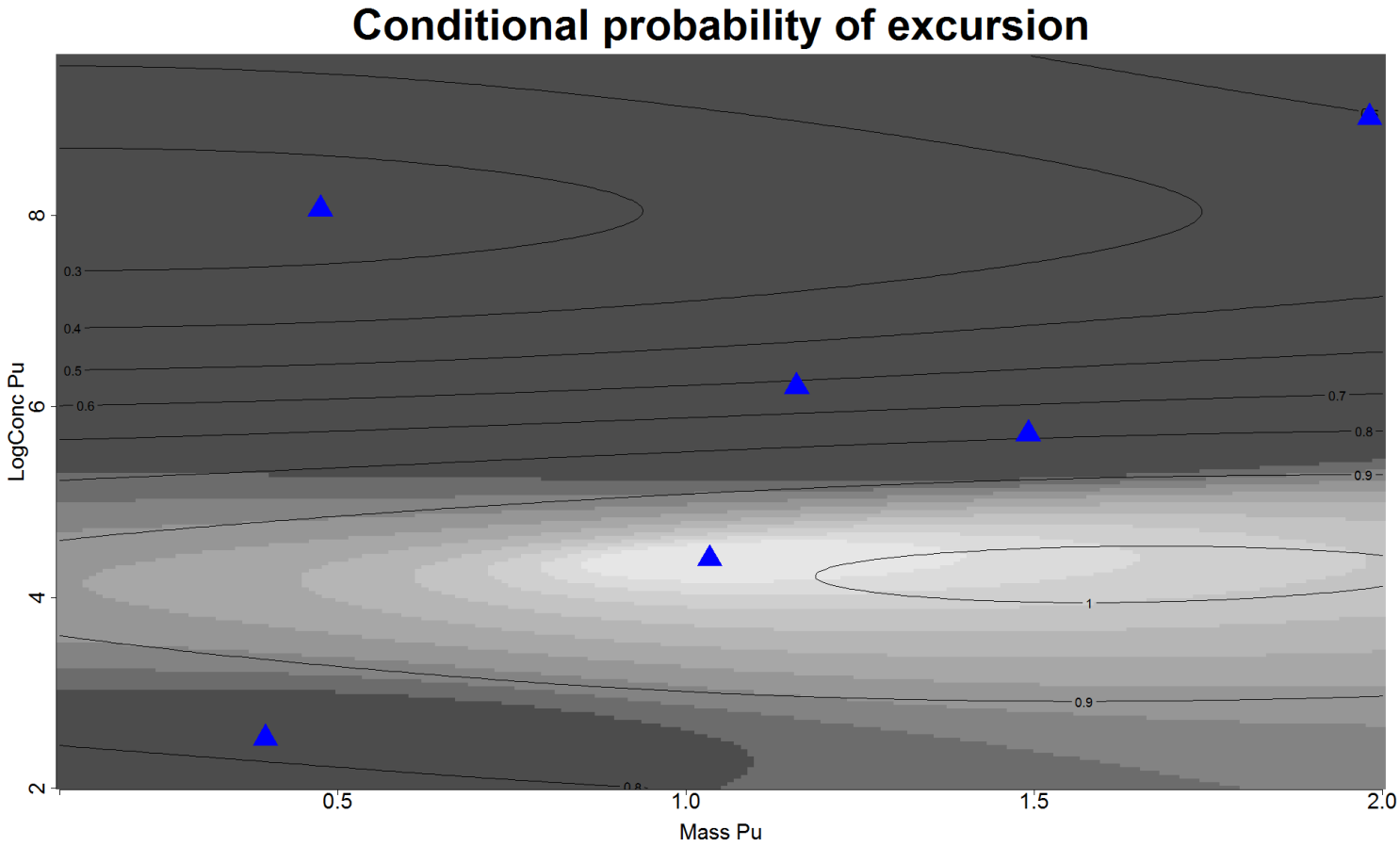


Kriging (Gaussian Process Interpolation)





$$\begin{cases} m_n(\mathbf{x}) = m(\mathbf{x}) + k(\mathbf{X}_n, \mathbf{x})^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} (f(\mathbf{X}_n) - m(\mathbf{X}_n)) \\ s_n^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - k(\mathbf{X}_n, \mathbf{x})^T k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x}) \end{cases}$$



$$p_n(\mathbf{x}) = P_n(\mathbf{x} \in \Gamma) = P_n(Z(\mathbf{x}) \geq t) = \Phi\left(\frac{m_n(\mathbf{x}) - t}{s_n(\mathbf{x})}\right)$$

SUR strategies for inversion and related problems

Let us focus first on estimating the measure of excursion

$$\alpha^\star := \mu(\Gamma^\star)$$

where μ is some prescribed finite measure on $(D, \mathcal{B}(D))$.

SUR strategies for inversion and related problems

Let us focus first on estimating the measure of excursion

$$\alpha^\star := \mu(\Gamma^\star)$$

where μ is some prescribed finite measure on $(D, \mathcal{B}(D))$.

Defining $\alpha := \mu(\Gamma)$, a number of quantities involving the distribution of α conditional on $Z(\mathbf{x}_1), \ldots Z(\mathbf{x}_n)$ can be calculated (in particular moments).

SUR strategies for inversion and related problems

Let us focus first on estimating the measure of excursion

$$\alpha^\star := \mu(\Gamma^\star)$$

where μ is some prescribed finite measure on $(D, \mathcal{B}(D))$.

Defining $\alpha := \mu(\Gamma)$, a number of quantities involving the distribution of α conditional on $Z(\mathbf{x}_1), \ldots Z(\mathbf{x}_n)$ can be calculated (in particular moments).

Approach considered here: sequentially reducing the excursion volume variance thanks to Stepwise Uncertainty Reduction (SUR) strategies

SUR strategies for inversion and related problems

We consider 1-step-lookahead optimal SUR strategies:

- Define a notion of uncertainty at time n: $H_n \geq 0$ (e.g., $var_n(\alpha)$).
- Reduce this uncertainty by evaluating Z at new points
- Sequential settings: evaluate sequentially the location $\mathbf{x}_{n+1}^{\star}$ minimizing the so-called SUR criterion associated with H_n:

$$J_n(\mathbf{x}_{n+1}) := \mathbb{E}_n(H_{n+1}(\mathbf{x}_{n+1}))$$

SUR strategies for inversion and related problems

We consider 1-step-lookahead optimal SUR strategies:

- Define a notion of uncertainty at time n: $H_n \geq 0$ (e.g., $var_n(\alpha)$).
- Reduce this uncertainty by evaluating Z at new points
- Sequential settings: evaluate sequentially the location $\mathbf{x}_{n+1}^{\star}$ minimizing the so-called SUR criterion associated with H_n:

$$J_n(\mathbf{x}_{n+1}) := \mathbb{E}_n(H_{n+1}(\mathbf{x}_{n+1}))$$

See notably the following paper and seminal references therein:

J. Bect, D. Ginsbourger, L. Li, V. Picheny and E. Vazquez.
Sequential design of computer experiments for the estimation of a probability of failure.
Statistics and Computing, 22(3):773-793, 2012.

SUR strategies: Two candidate uncertainties

Two possible definitions for the uncertainty H_n are considered here:

$$H_n := Var_n(\alpha)$$

$$\widetilde{H}_n := \int_D p_n(1-p_n)\mathrm{d}\mu$$

SUR strategies: Two candidate uncertainties

Two possible definitions for the uncertainty H_n are considered here:

$$H_n := Var_n(\alpha)$$
$$\widetilde{H}_n := \int_D p_n(1-p_n)\mathrm{d}\mu$$

Uncertainties:

$$H_n := Var_n(\alpha)$$
$$\widetilde{H}_n := \int_{\mathbb{X}} p_n(1-p_n)\mathrm{d}\mu$$

SUR criteria:

$$J_n(\mathbf{x}) := \mathbb{E}_n(Var_{n+1}(\alpha))$$
$$\widetilde{J}_n(\mathbf{x}) := \mathbb{E}_n\left(\int_D p_{n+1}(1-p_{n+1})\mathrm{d}\mu\right)$$

Main challenge to calculate $\widetilde{J}_n(\mathbf{x})$ (similar for $J_n(\mathbf{x})$): Obtain a closed form expression for $\mathbb{E}_n\left(p_{n+1}(1-p_{n+1})\right)$ and integrate it.

Deriving SUR criteria

Proposition

$$\mathbb{E}_n(p_{n+1}(\mathbf{x})(1-p_{n+1}(\mathbf{x}))) = \Phi_2\left(\begin{pmatrix} a(\mathbf{x}) \\ -a(\mathbf{x}) \end{pmatrix}, \begin{pmatrix} c(\mathbf{x}) & 1-c(\mathbf{x}) \\ 1-c(\mathbf{x}) & c(\mathbf{x}) \end{pmatrix}\right)$$

- $\Phi_2(\cdot, M)$*: c.d.f. of centred bivariate Gaussian with covariance matrix* M
- $a(\mathbf{x}) := (m_n(\mathbf{x}) - t)/s_{n+q}(\mathbf{x})$,
- $c(\mathbf{x}) := s_n^2(\mathbf{x})/s_{n+q}^2(\mathbf{x})$

C. Chevalier, J. Bect, D. Ginsbourger, V. Picheny, E. Vazquez and Y. Richet.
Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set.
Technometrics, 56(4):455-465, 2014.

Deriving SUR criteria

Proposition

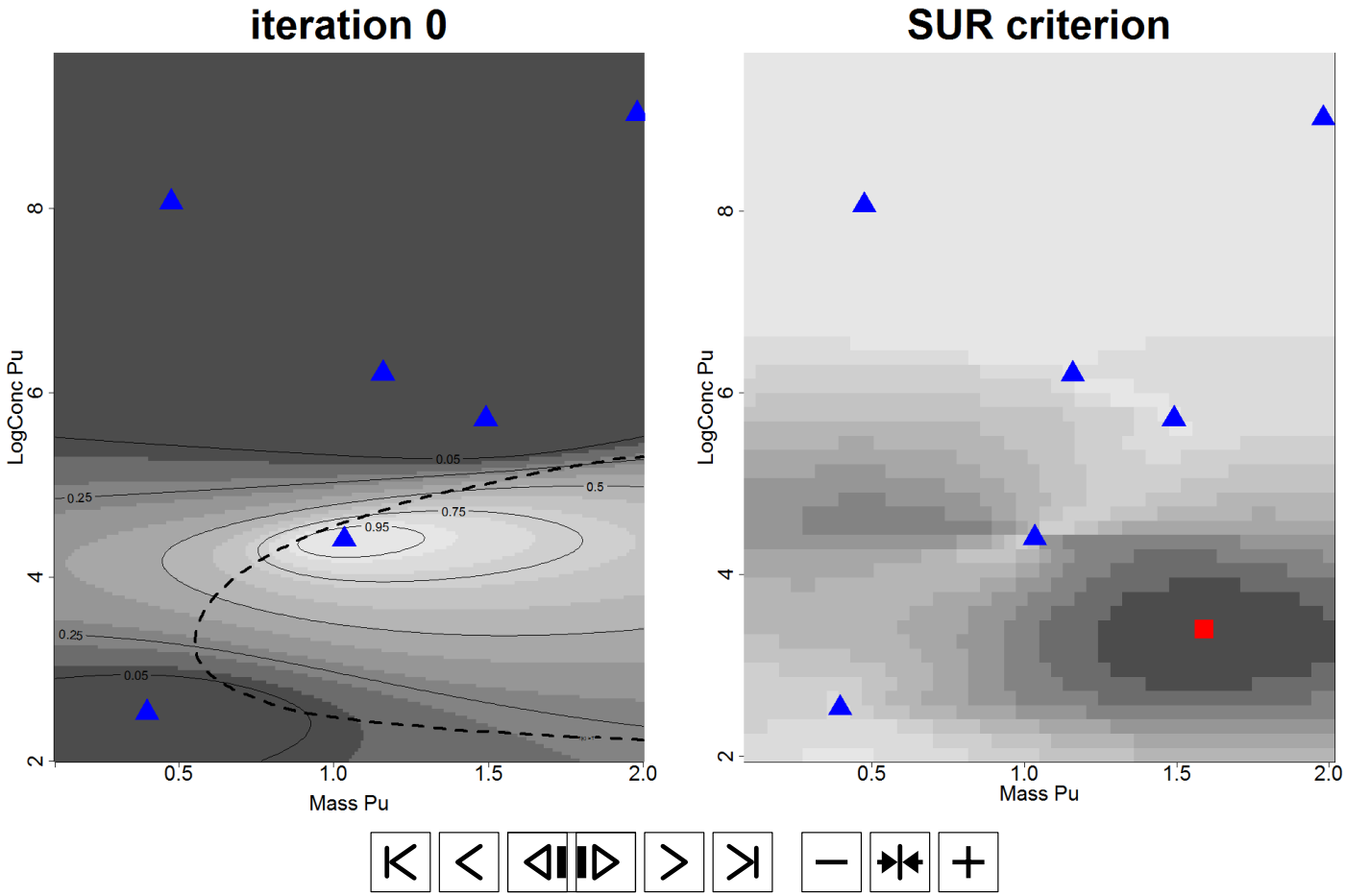
$$\mathbb{E}_n(p_{n+1}(\mathbf{x})(1-p_{n+1}(\mathbf{x}))) = \Phi_2\left(\begin{pmatrix} a(\mathbf{x}) \\ -a(\mathbf{x}) \end{pmatrix}, \begin{pmatrix} c(\mathbf{x}) & 1-c(\mathbf{x}) \\ 1-c(\mathbf{x}) & c(\mathbf{x}) \end{pmatrix}\right)$$

- $\Phi_2(\cdot, M)$*: c.d.f. of centred bivariate Gaussian with covariance matrix* M
- $a(\mathbf{x}) := (m_n(\mathbf{x}) - t)/s_{n+q}(\mathbf{x})$,
- $c(\mathbf{x}) := s_n^2(\mathbf{x})/s_{n+q}^2(\mathbf{x})$


C. Chevalier, J. Bect, D. Ginsbourger, V. Picheny, E. Vazquez and Y. Richet.
Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set.
Technometrics, 56(4):455-465, 2014.


C. Chevalier, V. Picheny and D. Ginsbourger.
The KrigInv package: An efficient and user-friendly R implementation of Kriging-based inversion algorithms.
Computational Statistics & Data Analysis, 71:1021-1034, 2014

Back to the test case with SUR batch



Further questions about SUR and UQ on sets

About the consistency:

J. Bect, F. Bachoc and D. Ginsbourger (2016).

A supermartingale approach to Gaussian process based sequential design of experiments.

HAL/Arxiv paper (hal-01351088, Arxiv: 1608.01118).

Further questions about SUR and UQ on sets

About the consistency:

J. Bect, F. Bachoc and D. Ginsbourger (2016).

A supermartingale approach to Gaussian process based sequential design of experiments.

HAL/Arxiv paper (hal-01351088, Arxiv: 1608.01118).

Abut conditional excursion set simulation:

D. Azzimonti, J. Bect, C. Chevalier and D. Ginsbourger (2016).

Quantifying uncertainties on excursion sets under a Gaussian random field prior.

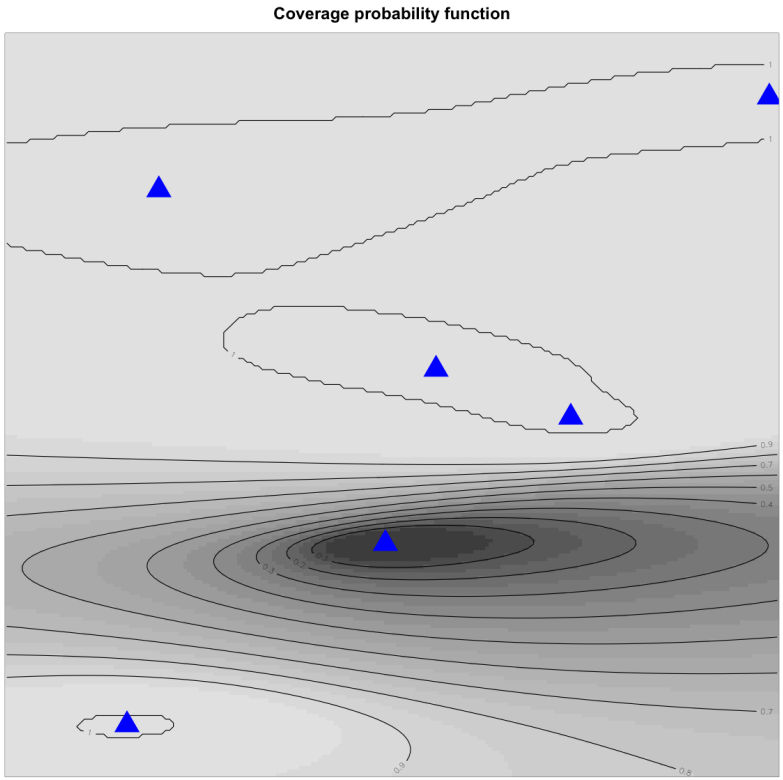
SIAM/ASA Journal on Uncertainty Quantification.

How to summarize the posterior distribution of sets?

Let us now reverse the perspective and focus on excursions below t, i.e. with $\Gamma = \{\mathbf{x} \in D : Z(\mathbf{x}) \leq t\}$ and $p_n : \mathbf{x} \in D \to p_n(x) = P_n(Z(\mathbf{x}) \leq t)$

How to summarize the posterior distribution of sets?

Let us now reverse the perspective and focus on excursions below t, i.e. with $\Gamma = \{\mathbf{x} \in D : Z(\mathbf{x}) \leq t\}$ and $p_n : \mathbf{x} \in D \to p_n(x) = P_n(Z(\mathbf{x}) \leq t)$



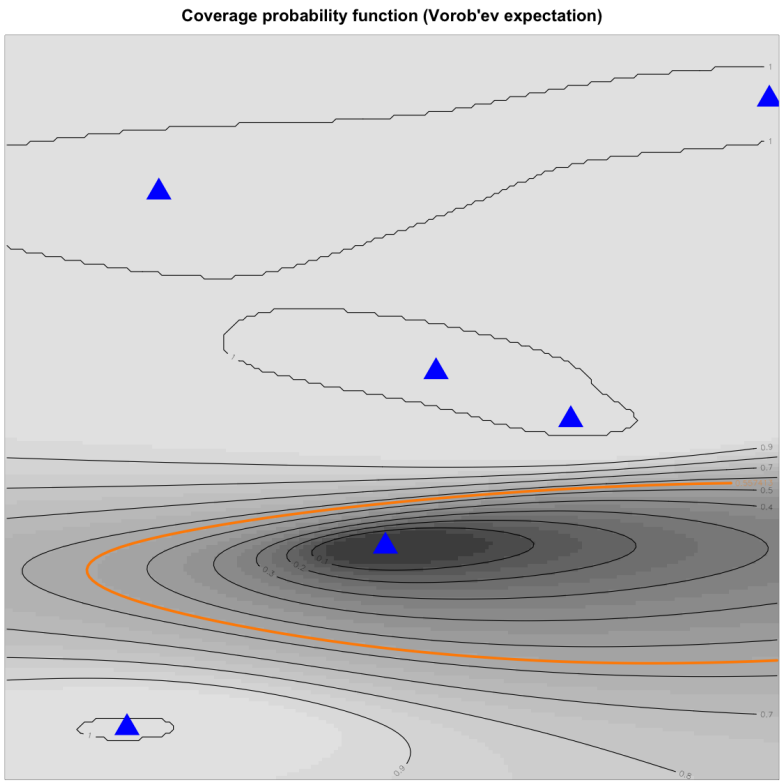
Define the (conditional) quantiles of Γ as $\rho-$level sets of p_n:

$$\begin{aligned} Q_\rho &:= \{\mathbf{x} \in D : p_n(\mathbf{x}) \geq \rho\} \\ &= \{\mathbf{x} \in D : P_n(Z(\mathbf{x}) \leq t) \geq \rho\}. \end{aligned}$$

How well Q_ρ estimates Γ can be quantified for instance through the "expected deviation":

$$\mathbb{E}_n\left(\mu(Q_\rho \Delta \Gamma)\right)$$

Estimates of $\Gamma^\star$: the Vorob'ev expectation



The **Vorob'ev expectation of** $\Gamma \mid (Z_{x_1} = f(x_1), \ldots, Z_{x_n} = f(x_n))$ is the $\rho^\star$ level set of p_n such that

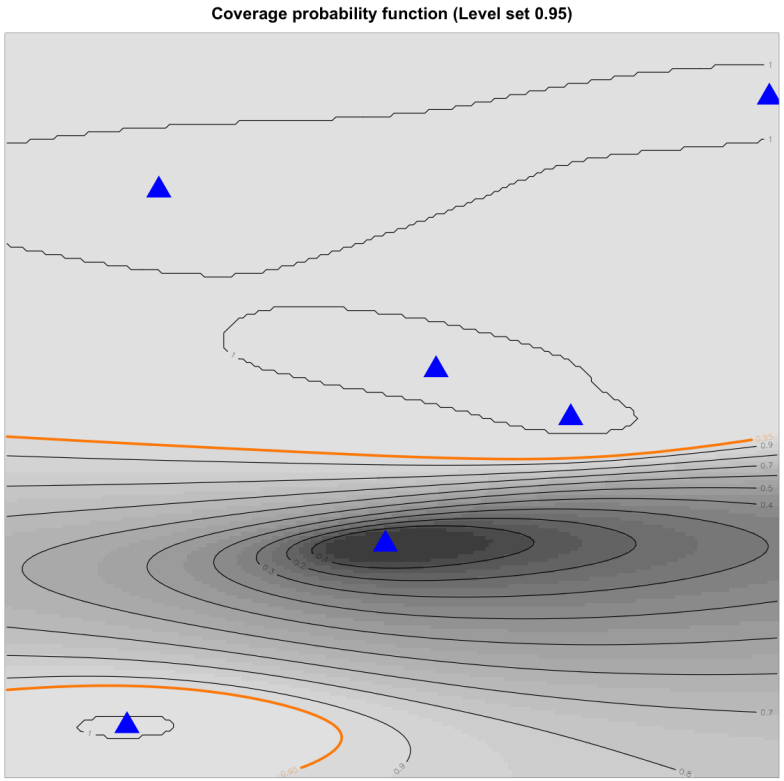
$$\mu(Q_{\rho^\star}) = \mathbb{E}_n[\mu(\Gamma)].$$

It is a state of the art result that $Q_{\rho^\star}$ minimizes $S \to \mathbb{E}_n\left(\mu(S \Delta \Gamma)\right)$ among all closed sets $S \subset \mathbb{R}^d$ with volume $\mathbb{E}_n[\mu(\Gamma)]$.

C. Chevalier, D. Ginsbourger, J. Bect, and Molchanov, I.
Estimating and quantifying uncertainties on level sets using the Vorob'ev expectation and deviation with Gaussian process models.
mODa 10 Advances in Model-Oriented Design and Analysis, Physica-Verlag HD, 2013.

Estimates of $\Gamma^\star$: some limitations of Q_ρ quantiles

In practice one often wish to give **confidence statements** on the estimates.

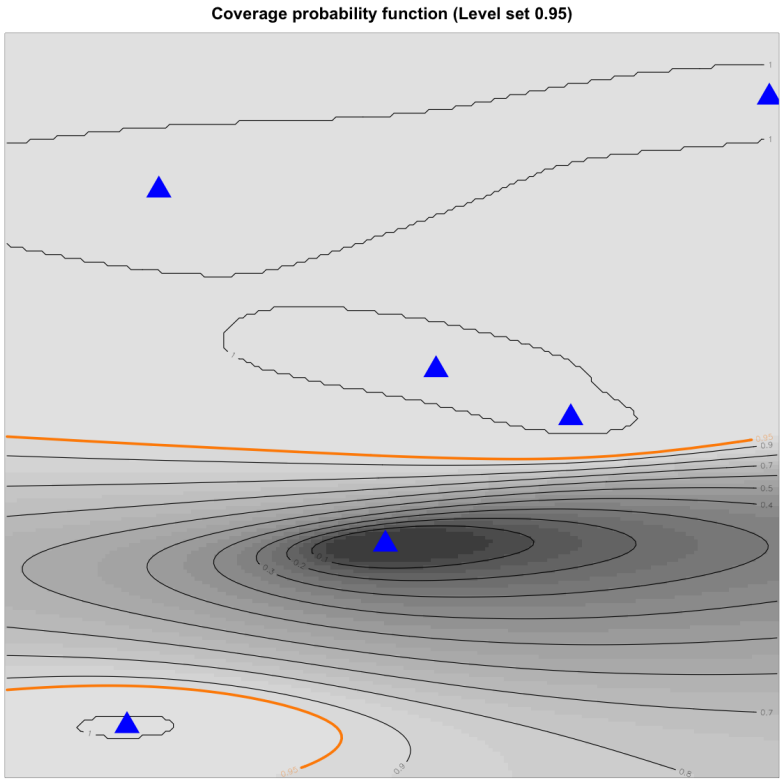


Q_ρ contains points which have marginal probability at least ρ of being in Γ.

$\Rightarrow$ no confidence statement on the probability of the actual excursion set containing this specific estimate.

Estimates of $\Gamma^\star$: some limitations of Q_ρ quantiles

In practice one often wish to give **confidence statements** on the estimates.



Q_ρ contains points which have marginal probability at least ρ of being in Γ.

$\Rightarrow$ no confidence statement on the probability of the actual excursion set containing this specific estimate.

E.g., the probabilities of Q_ρ containing the excursion set (computed on a grid) are

- 0.67 for $\rho = 0.95$
- 0.009 for $\rho = 0.5$
- 0.019 for $\rho = 0.56$ (Vorob'ev)

Conservative Estimates of $\Gamma^\star$

We denote by **conservative estimate** for $\Gamma \mid (Z_{x_1} = f(x_1), \ldots, Z_{x_n} = f(x_n))$ at level β the largest Q_ρ such that $P_n(Q_\rho \subset \Gamma) \geq \beta$:

$$E_{t,\alpha} = \arg\max_{Q_\rho} \{|Q_\rho| : P_n(Q_\rho \subset \{Z_x \leq t\}) \geq \beta\}$$

D. Bolin, F. Lindgren.
Excursion and contour uncertainty regions for latent Gaussian models.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2014.

Conservative Estimates of $\Gamma^\star$

We denote by **conservative estimate** for $\Gamma \mid (Z_{x_1} = f(x_1), \ldots, Z_{x_n} = f(x_n))$ at level β the largest Q_ρ such that $P_n(Q_\rho \subset \Gamma) \geq \beta$:

$$E_{t,\alpha} = \arg\max_{Q_\rho} \{|Q_\rho| : P_n(Q_\rho \subset \{Z_x \leq t\}) \geq \beta\}$$

D. Bolin, F. Lindgren.
Excursion and contour uncertainty regions for latent Gaussian models.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2014.

Such conservative estimate $E_{t,\beta}$ is hence

- the largest quantile such that, with probability β, the response is below the threshold simultaneously at each of its locations.
- based on a confidence statement on the whole set

Computing conservative estimates

The computation of a conservative estimate

$$E_{t,\beta} = \arg\max_{Q_\rho} \{|Q_\rho| : P_n(Q_\rho \subset \{Z_x \leq t\}) \geq \beta\}$$

presents two computational bottlenecks:

1. find the set with the maximum volume;
2. compute $P_n(Q_\rho \subset \{Z_x \leq t\})$.

Computing conservative estimates

The computation of a conservative estimate

$$E_{t,\beta} = \arg\max_{Q_\rho} \{|Q_\rho| : P_n(Q_\rho \subset \{Z_x \leq t\}) \geq \beta\}$$

presents two computational bottlenecks:

1. find the set with the maximum volume;
2. compute $P_n(Q_\rho \subset \{Z_x \leq t\})$.

For recent work on computing the last term, see for instance

D. Azzimonti and D. Ginsbourger (2016+).
Estimating orthant probabilities of high dimensional Gaussian vectors with an application to set estimation.
Hal/Arviv paper.

Computing $P_n(Q_\rho \subset \{Z_x \leq t\})$

If Q_ρ is discretized over a grid $W = \{w_1, \ldots, w_m\}$, then

$$P_n(Q_\rho \subset \{Z_x \leq t\}) = P_n(Z_{w_1} \leq t, \ldots, Z_{w_m} \leq t) = 1 - P_n\left(\max_{i=1,\ldots,m} Z_{w_i} > t\right)$$

Computing $P_n(Q_\rho \subset \{Z_x \leq t\})$

If Q_ρ is discretized over a grid $W = \{w_1, \ldots, w_m\}$, then

$$P_n(Q_\rho \subset \{Z_x \leq t\}) = P_n(Z_{w_1} \leq t, \ldots, Z_{w_m} \leq t) = 1 - P_n\left(\max_{i=1,\ldots,m} Z_{w_i} > t\right)$$

There exists a number of algorithms to estimate $P_n(Z_{w_1} \leq t, \ldots, Z_{w_m} \leq t)$:

1. deterministic QMC integration techniques: **Genz quadrature**
   - very fast and reliable in small dimensions;
   - not usable for dimensions higher than 1000.
2. pure **MC techniques**:
   - dimension independent;
   - high number of simulations for small variance.

Computing $P_n(Q_\rho \subset \{Z_x \leq t\})$

If Q_ρ is discretized over a grid $W = \{w_1, \ldots, w_m\}$, then

$$P_n(Q_\rho \subset \{Z_x \leq t\}) = P_n(Z_{w_1} \leq t, \ldots, Z_{w_m} \leq t) = 1 - P_n\left(\max_{i=1,\ldots,m} Z_{w_i} > t\right)$$

There exists a number of algorithms to estimate $P_n(Z_{w_1} \leq t, \ldots, Z_{w_m} \leq t)$:

1. deterministic QMC integration techniques: **Genz quadrature**
   - very fast and reliable in small dimensions;
   - not usable for dimensions higher than 1000.
2. pure **MC techniques**:
   - dimension independent;
   - high number of simulations for small variance.

IRSN test case

- an estimate with a good resolution requires an 100×100 grid for D;
- W is a grid with $+1000$ points for some Q_ρ, quadrature is hardly usable.

$P_n(\max_{w \in W} Z_w > T)$: proposed hybrid algorithm

Algorithm:

1. select q grid points, denoted $W_q \subset W$;
2. compute $p' = P(\max_{w \in W_q} Z_w > t)$ with Genz quadrature;
3. estimate $P_n(\max_{w \in W} Z_w > t)$ with

$$\hat{p} = p' + (1 - p')\hat{R}_q$$

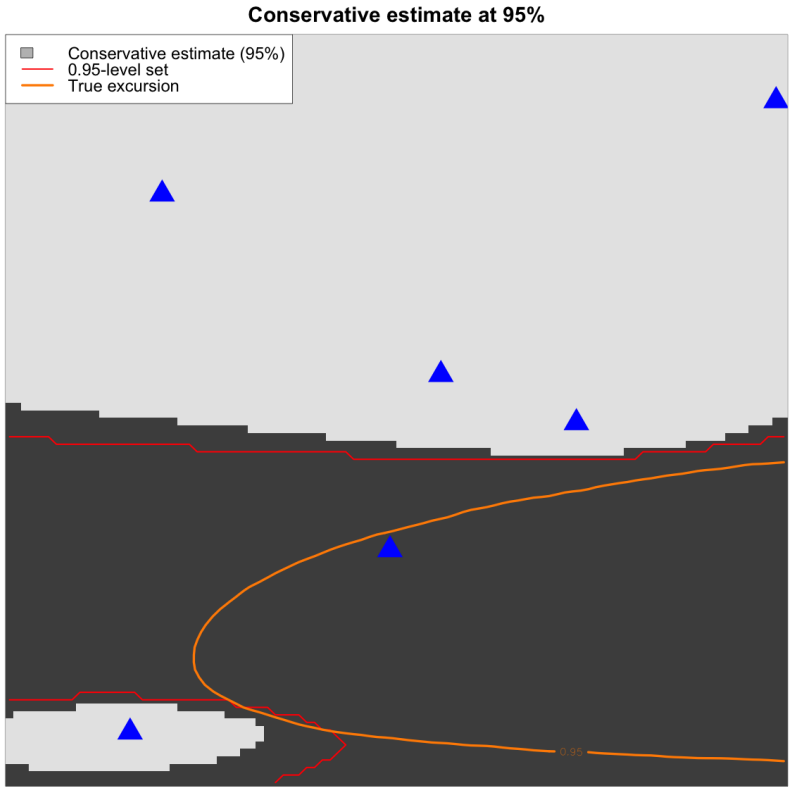
where $\hat{R}_q$ is a MC estimate of

$$R_q = P_n\left(\max_{w \in W \setminus W_q} Z_w > t \mid \max_{w \in W_q} Z_w \leq t\right)$$

Note: $P_n(\max_{w \in W_q} Z_w > t) = p' \leq p = P_n(\max_{w \in W} Z_w > t)$,

Computing the conservative estimate: test case

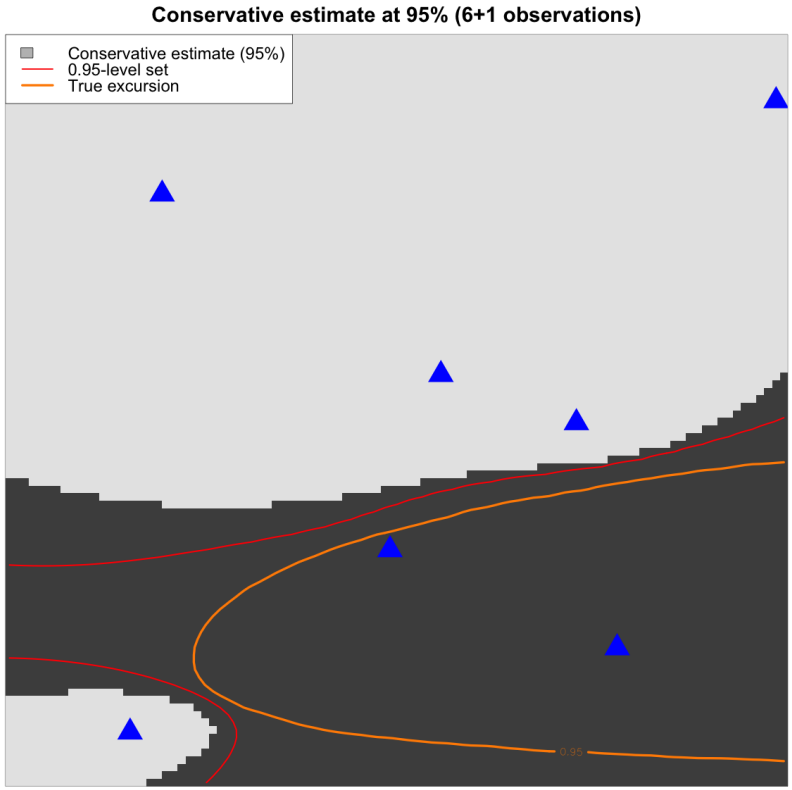
Our hybrid algorithm allowed us computing conservative estimates.



Conservative estimate at 95%

Computing the conservative estimate: test case

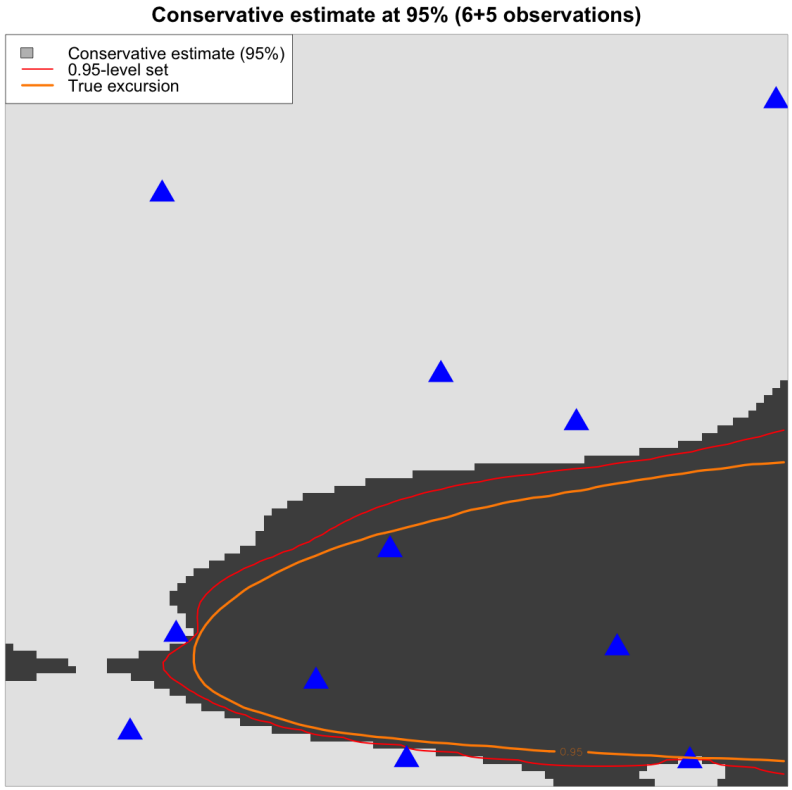
Now with 1 additional points obtained by SUR strategy . . .



Conservative estimate at 95%

Computing the conservative estimate: test case

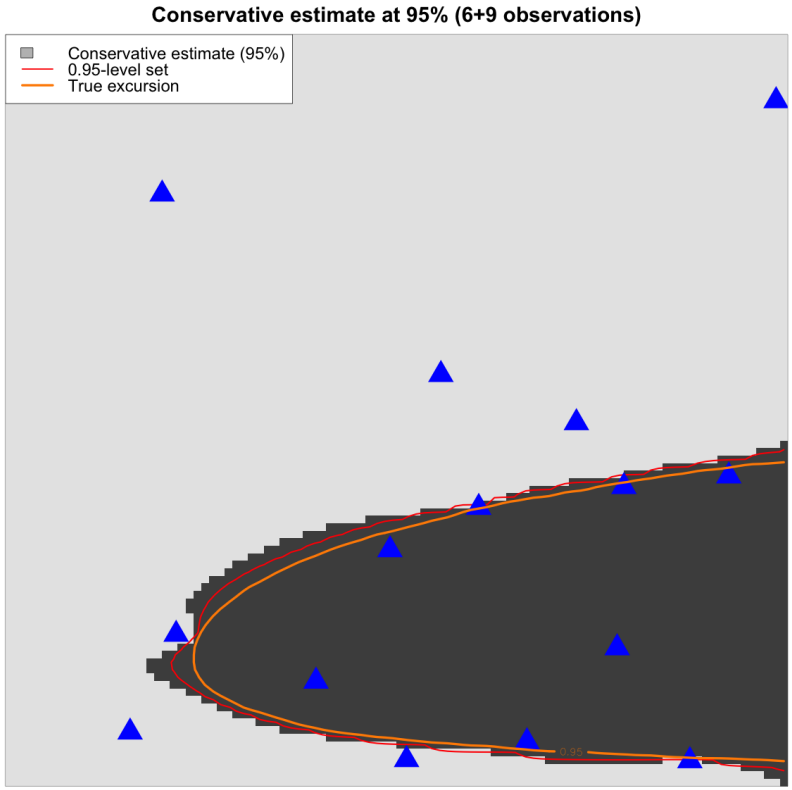
. . . now with 5 additional points from the SUR strategy. . .



Conservative estimate at 95%

Computing the conservative estimate: test case

. . . and finally with a total of 9 additional SUR points!



Conservative estimate at 95%

Perspectives

- Further improve the hybrid MC/QMC scheme $\Rightarrow$ notably within Dario Azzimonti's PhD work; Cf. his talk!.
- Transpose this workflow to other families of implicitly defined regions (ongoing).
- Consider families of set estimates beyond quantiles (ongoing).
- Derive further convergence properties for SUR strategies (ongoing).

Perspectives

- Further improve the hybrid MC/QMC scheme $\Rightarrow$ notably within Dario Azzimonti's PhD work; Cf. his talk!.
- Transpose this workflow to other families of implicitly defined regions (ongoing).
- Consider families of set estimates beyond quantiles (ongoing).
- Derive further convergence properties for SUR strategies (ongoing).

Acknowledgements: Drs Yann Richet and Grégory Caplin (French Nuclear Safety Institute) for providing the criticality safety test case.

Part IV

Incorporation of degeneracies and invariances in GP models

Proposition

Let Z be a measurable random field with path 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

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

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:

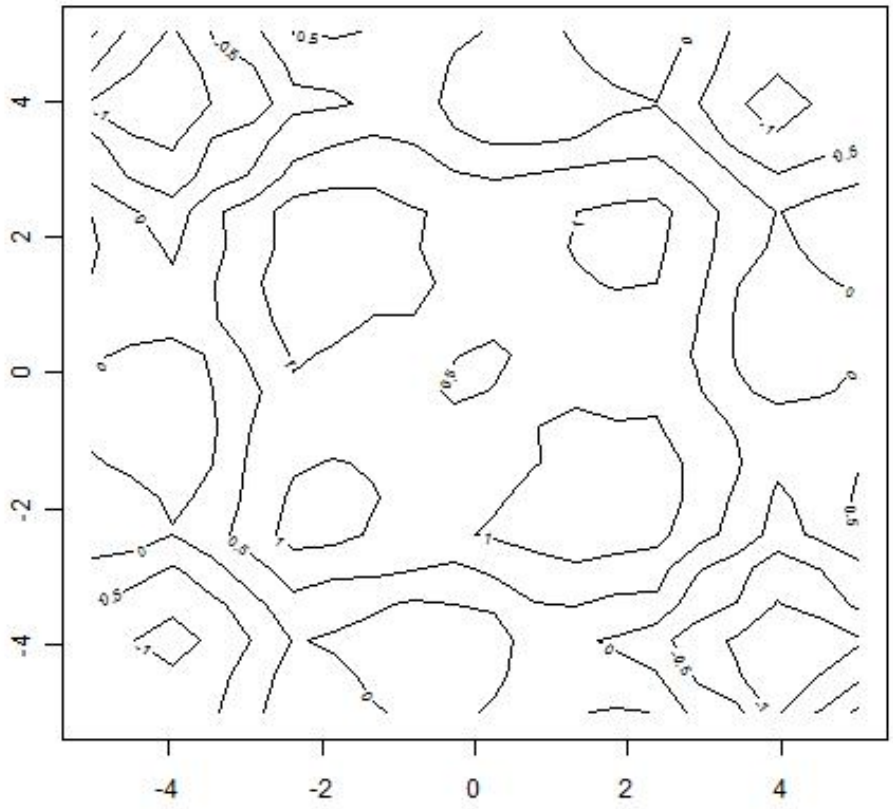
a) $\forall \mathbf{x} \in D \; \mathbb{P}(T(Z)_{\mathbf{x}} = 0) = 1$ *("$T(Z) = \mathbf{0}$ up to a modification")*

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.") .*

Invariance under the action of a finite group



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

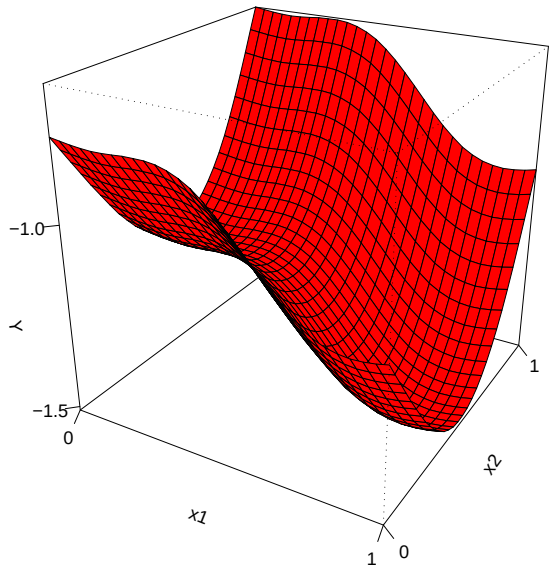
Let $D = \prod_i^d D_i$ where $D_i \subset \mathbb{R}$. $f \in \mathbb{R}^D$ is called additive when there exists $f_i \in \mathbb{R}^{D_i}$ $(1 \leq i \leq d)$ such that $f(\mathbf{x}) = \sum_{i=1}^d f_i(x_i)$ $(\mathbf{x} = (x_1, \ldots, x_d) \in D)$.

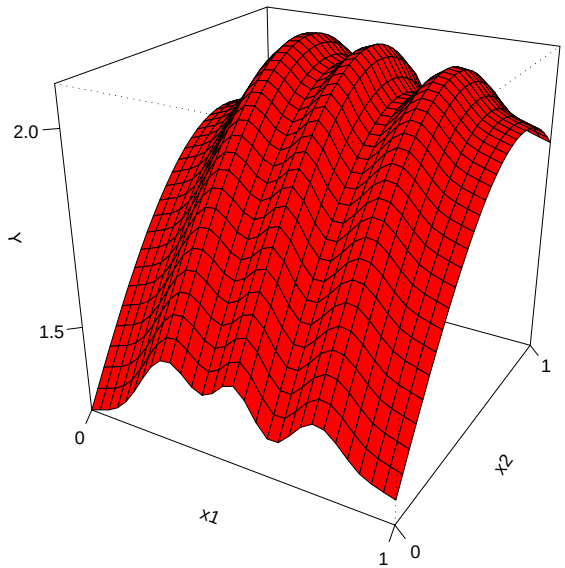
GRF models possessing additive paths (with $k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^d k_i(x_i, x_i')$) have been considered in Nicolas Durrande's Ph.D. thesis (2011):

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 $f(\mathbf{x}) = \sum_{i=1}^d f_i(x_i)$ $(\mathbf{x} = (x_1, \ldots, x_d) \in D)$.

GRF 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/related references

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

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

D. Ginsbourger, 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

Extension to further operators in the Gaussian case

In the Gaussian case, the last results can be extended to a wider class of operators using the Loève isometry Ψ between $\mathcal{L}(Z)$ (The Hilbert space generated by Z) and the RKHS associated with k, $\mathcal{H}(k)$.

Extension to further operators in the Gaussian case

In the Gaussian case, the last results can be extended to a wider class of operators using the Loève isometry Ψ between $\mathcal{L}(Z)$ (The Hilbert space generated by Z) and the RKHS associated with k, $\mathcal{H}(k)$.

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 $\mathcal{T} : \mathcal{H} \to \mathbb{R}^D$ satisfying

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

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

Examples (Gaussian case)

a) Let ν be a measure on D s.t. $\int_D \sqrt{k(\mathbf{u},\mathbf{u})} d\nu(\mathbf{u}) < +\infty$.
Then Z 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

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

satisfies the above condition.

Examples (Gaussian case)

a) Let ν be a measure on D s.t. $\int_D \sqrt{k(\mathbf{u},\mathbf{u})} d\nu(\mathbf{u}) < +\infty$.
Then Z 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

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

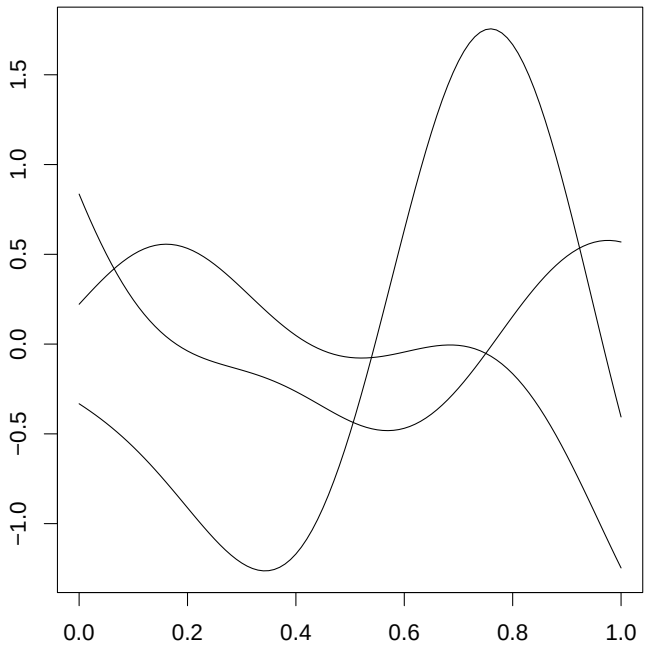
satisfies the above condition.

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

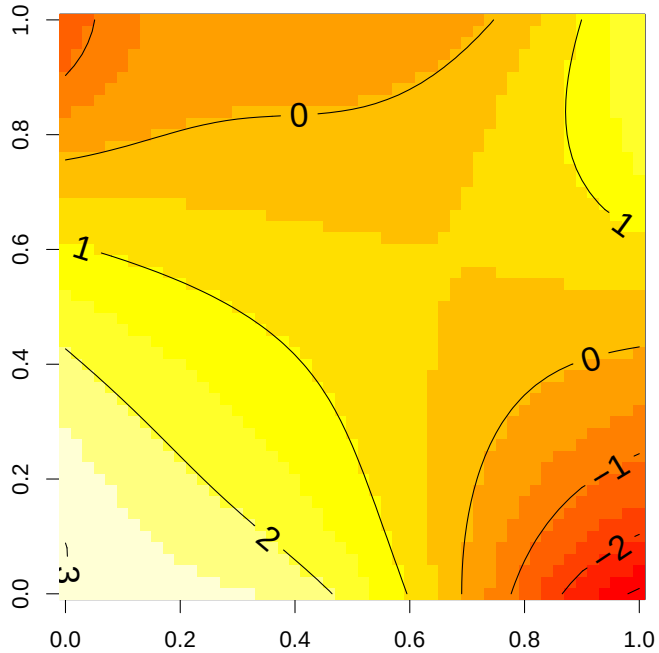
An example of such harmonic kernel over $\mathbb{R}^2 \times \mathbb{R}^2$ can be found in the recent literature (Schaback et al. 2009):

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

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

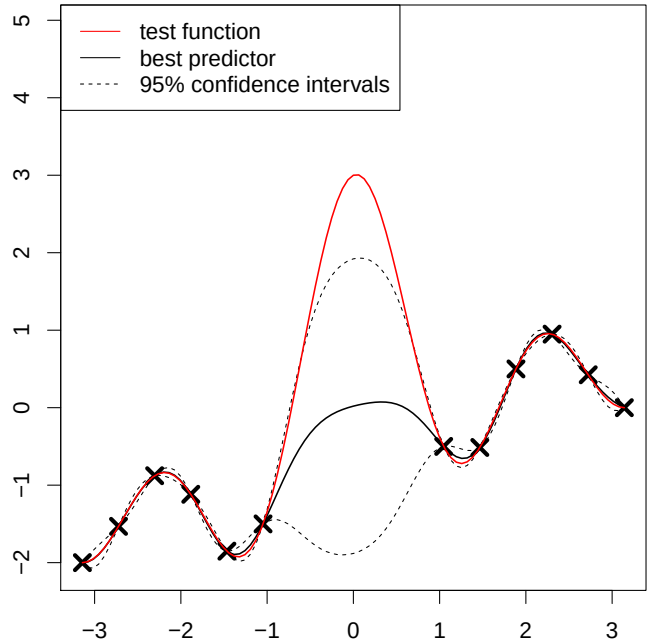
Proposition

Let

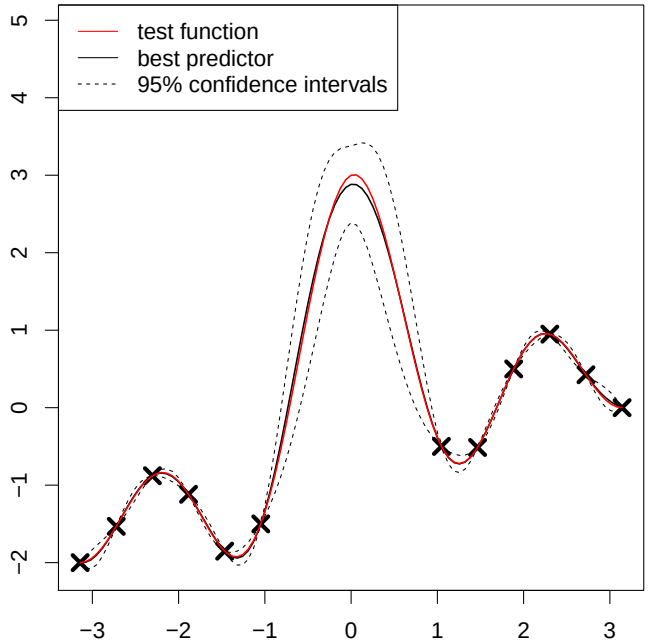
- $\mathcal{F}, \mathcal{G}$ *be real separable Banach spaces,*
- μ *be a Gaussian measure on* $\mathcal{B}(\mathcal{F})$ *with mean zero and covariance operator* C_μ
- $T : \mathcal{F} \longrightarrow \mathcal{F}$ *be a bounded linear operator such that* $TC_\mu T^\star = 0_{\mathcal{F}^\star \longrightarrow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

Numerical application: Kriging with a centred kernel



(c) GPR with kernel k

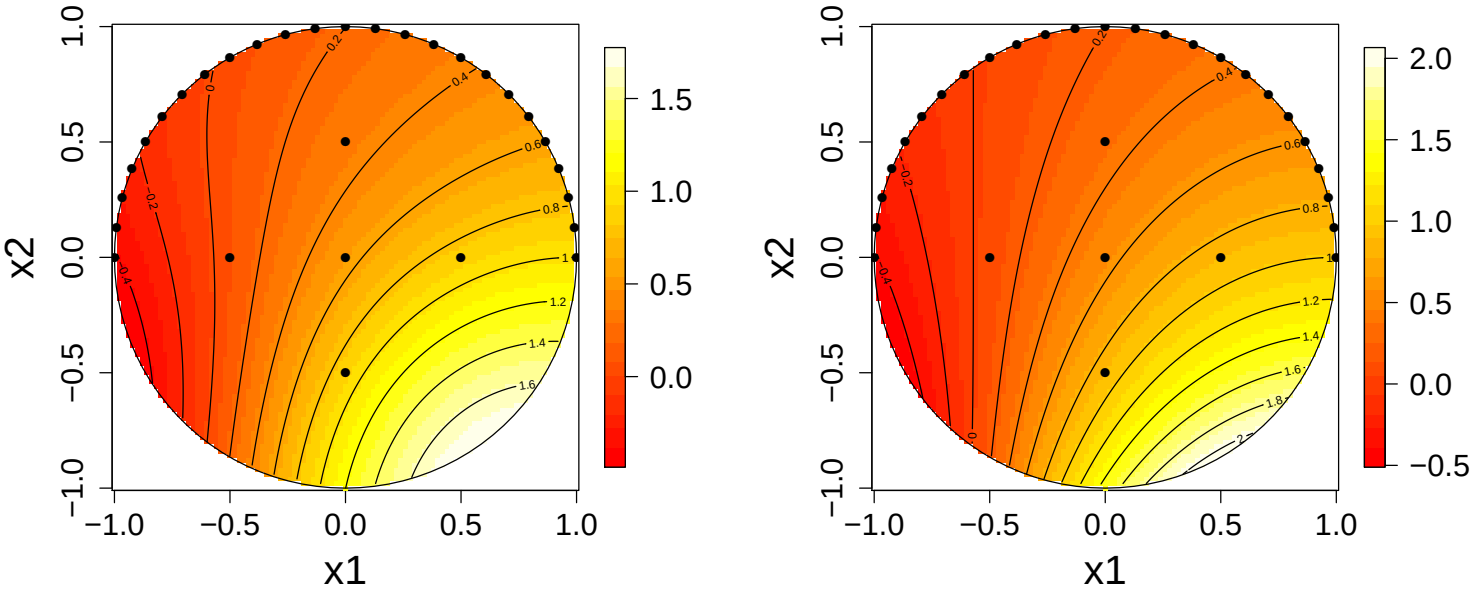


(d) GPR with kernel k_0

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

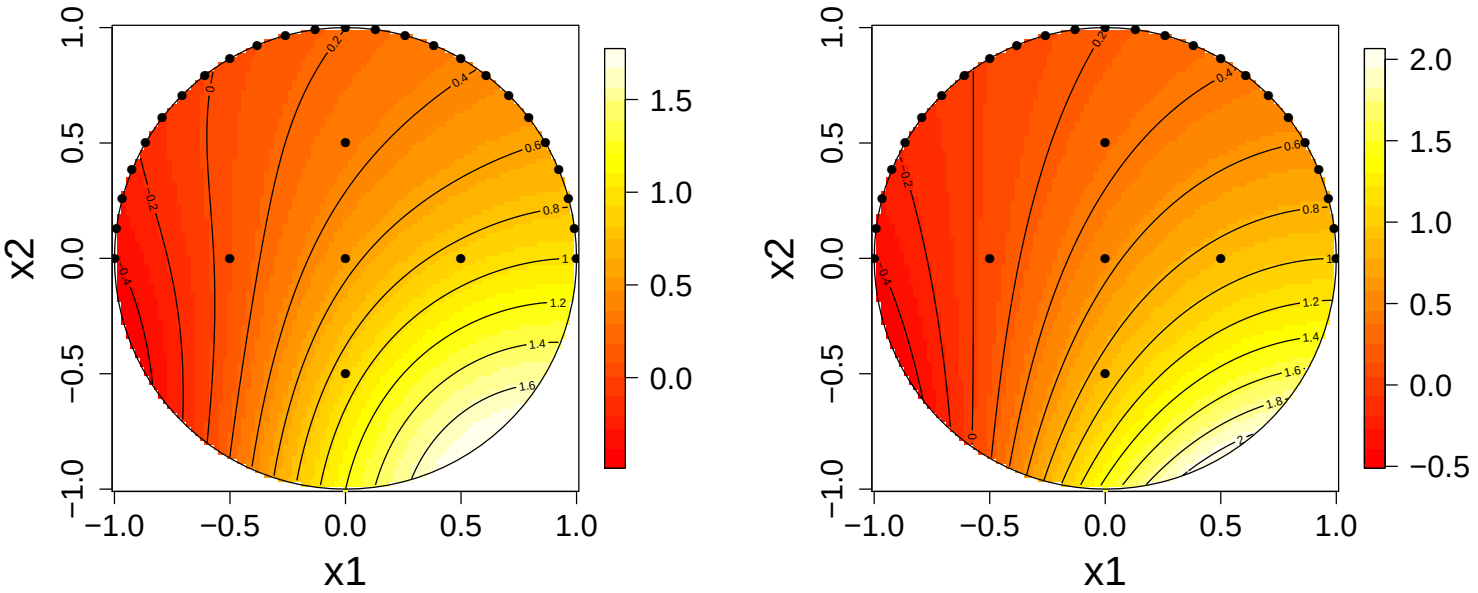
Numerical application bis: maximum of a harmonic *f*

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



Numerical application bis: maximum of a harmonic *f*

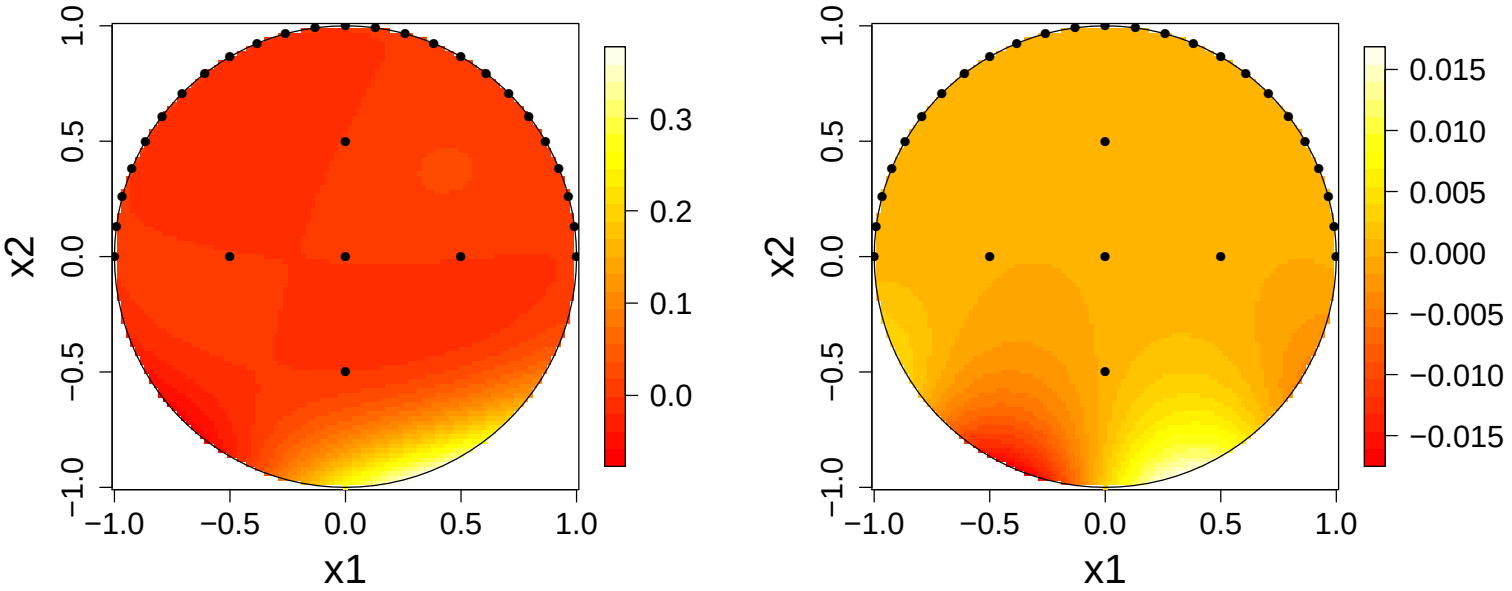
Here we consider approximating a harmonic function (left/right: Gaussian/harmonic kernels) and estimating its maximum by GRF modelling.



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)

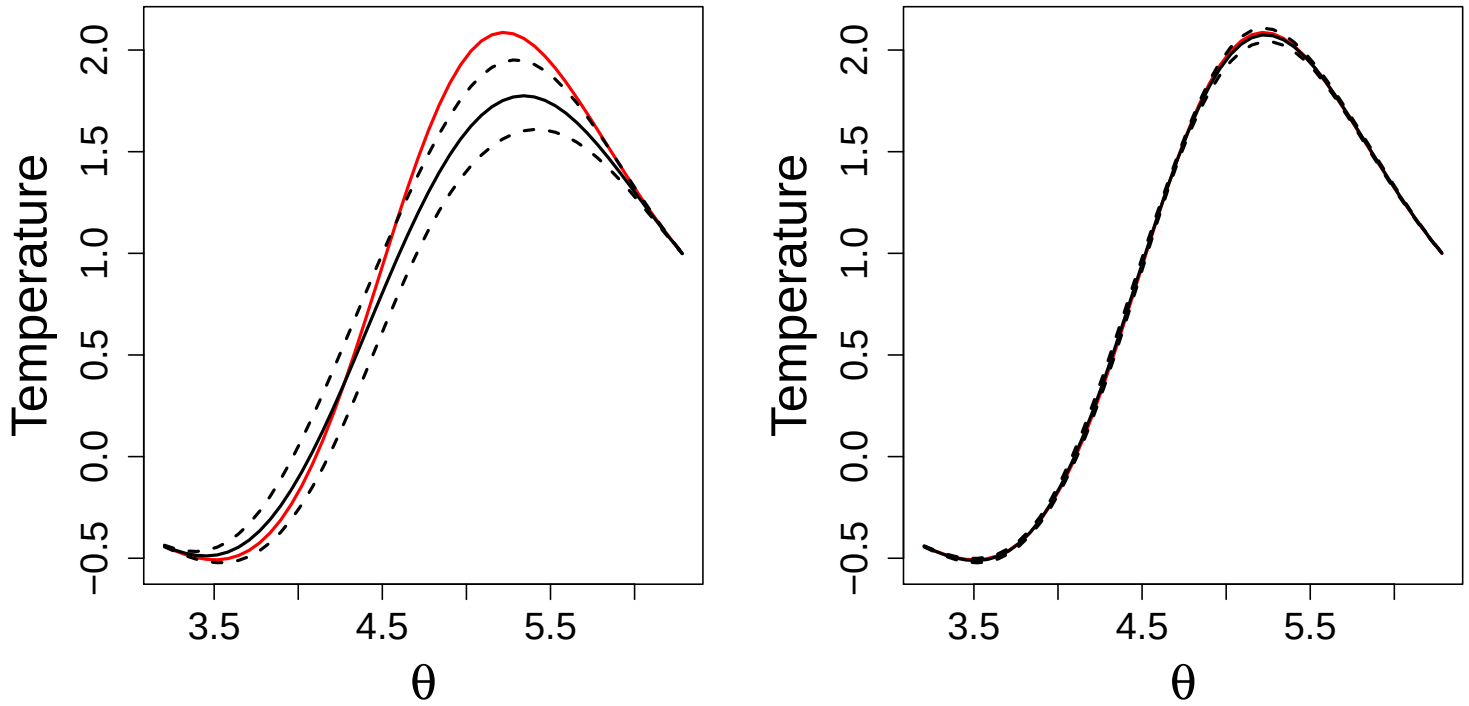
Numerical application bis: maximum of a harmonic *f*

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



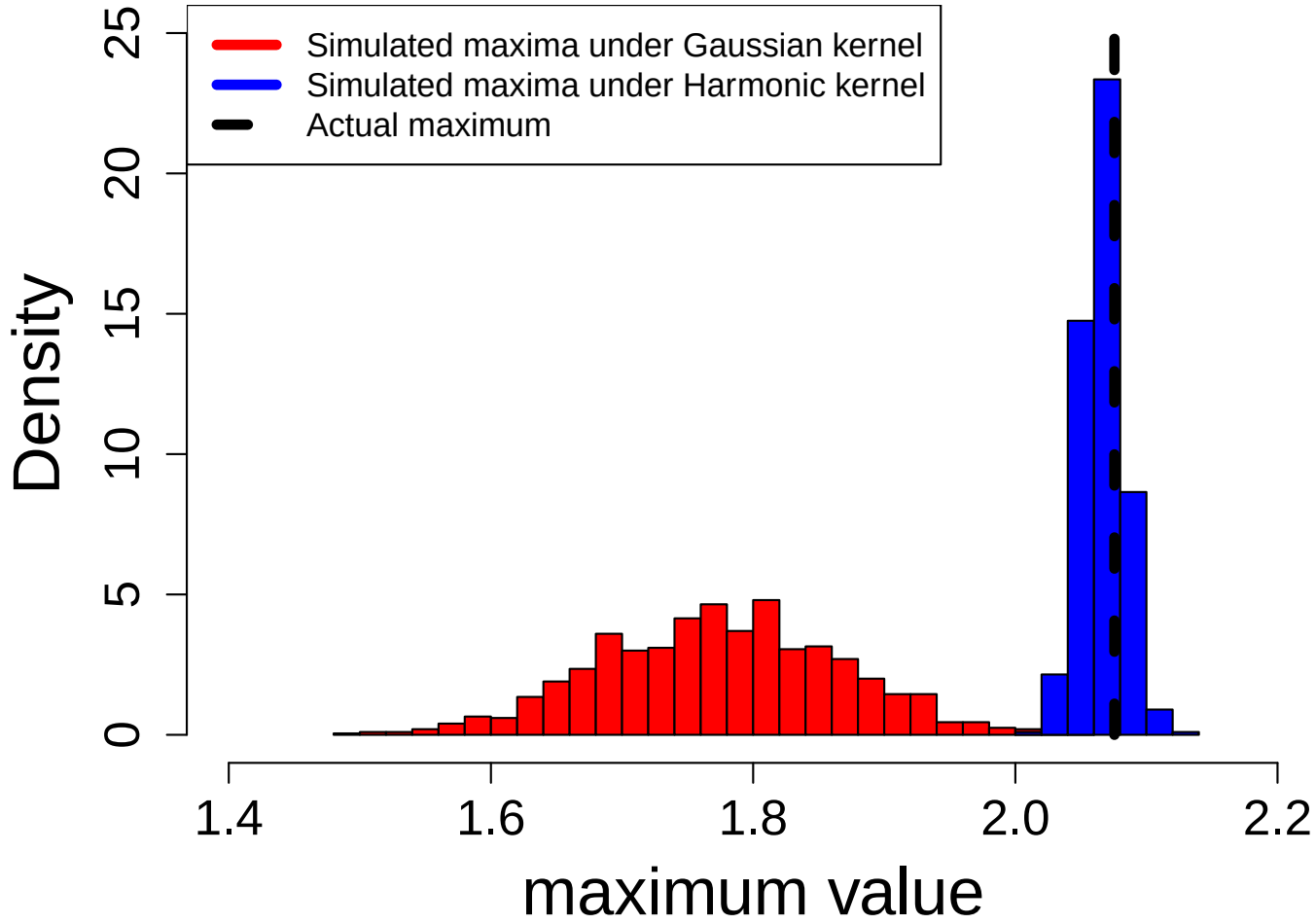
Numerical application bis: maximum of a harmonic *f*

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



Numerical application bis: maximum of a harmonic *f*

Conditional simulations of the maximum under the two GRF models.



Further references

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

D. Ginsbourger, 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

Further references


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


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


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. Ginsbourger, 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

Part V

Appendix

The No-Empty-Ball (NEB) property back

k has the NEB property if for all sequences $(\mathbf{x}_n)$ in D and $\boldsymbol{y} \in D$ the following assertions are equivalent:

- $\boldsymbol{y}$ is an adherent point of the set $\{\mathbf{x}_n, n \geq 1\}$
- $s_n^2(\boldsymbol{y}; \mathbf{x}_1, \ldots, \mathbf{x}_n) \to 0 \ (n \to \infty)$

where $s_n^2(\boldsymbol{y}; \mathbf{x}_1, \ldots, \mathbf{x}_n)$ stands for $s_n^2(\boldsymbol{y})$ when f is known at $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

E. Vazquez, J. Bect.

Convergence properties of the expected improvement algorithm with fixed mean and covariance functions.

Journal of Statistical Planning and Inference, 2010.

The No-Empty-Ball (NEB) property back

k has the NEB property if for all sequences $(\mathbf{x}_n)$ in D and $\boldsymbol{y} \in D$ the following assertions are equivalent:

- $\boldsymbol{y}$ is an adherent point of the set $\{\mathbf{x}_n, n \geq 1\}$
- $s_n^2(\boldsymbol{y}; \mathbf{x}_1, \ldots, \mathbf{x}_n) \to 0$ $(n \to \infty)$

where $s_n^2(\boldsymbol{y}; \mathbf{x}_1, \ldots, \mathbf{x}_n)$ stands for $s_n^2(\boldsymbol{y})$ when f is known at $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

E. Vazquez, J. Bect.

Convergence properties of the expected improvement algorithm with fixed mean and covariance functions.

Journal of Statistical Planning and Inference, 2010.

A proven sufficient condition for the NEB to hold is that k is stationary and possesses a spectral density S such that S^{-1} has at most polynomial growth.

Multipoint EI: closed form based on Tallis' formula

Denote $\mathbf{Y} = Z(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+q}) \sim \mathcal{N}_q(\boldsymbol{m}, \Sigma)$, and for $k \in \{1, \ldots, q\}$ consider $\mathbf{Z}^{(k)}$ defined by: $Z_j^{(k)} := Y_j - Y_k$ if $j \neq k$ and $Z_k^{(k)} := -Y_k$.

Let $\boldsymbol{m}^{(k)}$ and $\Sigma^{(k)}$ be the conditional mean and covariance matrix of $\mathbf{Z}^{(k)}$ at step n, and $\boldsymbol{b}^{(k)} \in \mathbb{R}^q$ be defined by $b_k^{(k)} = -T$ and $b_j^{(k)} = 0$ if $j \neq k$. Applying Tallis' formula yields

$$EI_n(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+q}) = \sum_{k=1}^{q} \left((m_k - T) p_k + \sum_{i=1}^{q} \Sigma_{ik}^{(k)} \varphi_{m_i^{(k)}, \Sigma_{ii}^{(k)}}(b_i^{(k)}) \Phi_{q-1}\left(\mathbf{c}_{.i}^{(k)}, \Sigma_{.i}^{(k)}\right) \right)$$

where:

- $p_k := \mathbb{P}(\mathbf{Z}^{(k)} \leq \boldsymbol{b}^{(k)}) = \Phi_q(\boldsymbol{b}^{(k)} - \boldsymbol{m}^{(k)}, \Sigma^{(k)})$.
- $\Phi_q(\boldsymbol{u}, \Sigma)$ ($\boldsymbol{u} \in \mathbb{R}^q, \Sigma \in \mathbb{R}^{q \times q}, q \geq 1$) is the c.d.f. of the centered multivariate Gaussian distribution with covariance matrix Σ.
- $\mathbf{c}_{.i}^{(k)}$ and $\Sigma_{.i}^{(k)}$ are the conditional mean and covariance matrix of random vector $(Z_1^{(k)}, \ldots, Z_{i-1}^{(k)}, Z_{i+1}^{(k)}, \ldots, Z_q^{(k)})$ knowing $Z_i^{(k)}$. Goto example

On regret bounds for the optimal strategy

Grünewälder et al. noticed than even if the optimal strategy is intractable, one can actually bound its expected regret by using a metric entropy bound.

Assumptions

1. For some $L_\mu \geq 0$, for any $\mathbf{x}, \mathbf{x}' \in D$, $|\mu(\mathbf{x}) - \mu(\mathbf{x}')| \leq L_\mu ||\mathbf{x} - \mathbf{x}'||_\infty$.
2. For some $L_k \geq 0$ and some $\alpha > 0$, for any $\mathbf{x}, \mathbf{x}' \in D$, $|k(\mathbf{x}, \mathbf{x}) - (\mathbf{x}, \mathbf{x}')| \leq L_k ||\mathbf{x} - \mathbf{x}'||_\infty^\alpha$.

On regret bounds for the optimal strategy

For any GRF satisfying the previous assumptions, and for $\mathbf{x}_1^\star, \ldots, \mathbf{x}_r^\star$ given by the optimal strategy,

$$\mathbb{E}\left[\sup_{\mathbf{x}\in D} Z_{\mathbf{x}} - \max(Z_{\mathbf{x}_1^\star}, \ldots, Z_{\mathbf{x}_r^\star})\right] \leq 4\sqrt{\frac{L_k \log(2r)}{(2\tilde{r})^\alpha}} + 15\sqrt{\frac{(\alpha+3)dL_k}{\alpha(2\tilde{r})^\alpha}} + \frac{L_\mu}{2\tilde{r}},$$

where $\tilde{r} = \left\lfloor r^{1/d} \right\rfloor$.

On regret bounds for the optimal strategy

For any GRF satisfying the previous assumptions, and for $\mathbf{x}_1^\star, \ldots, \mathbf{x}_r^\star$ given by the optimal strategy,

$$\mathbb{E}\left[\sup_{\mathbf{x}\in D} Z_{\mathbf{x}} - \max(Z_{\mathbf{x}_1^\star}, \ldots, Z_{\mathbf{x}_r^\star})\right] \leq 4\sqrt{\frac{L_k \log(2r)}{(2\tilde{r})^\alpha}} + 15\sqrt{\frac{(\alpha+3)dL_k}{\alpha(2\tilde{r})^\alpha}} + \frac{L_\mu}{2\tilde{r}},$$

where $\tilde{r} = \left\lfloor r^{1/d} \right\rfloor$. In addition, Grünewälder et al. showed that there exists a GRF Z satisfying the previous assumptions and such that

$$\mathbb{E}\left[\sup_{\mathbf{x}\in D} Z_{\mathbf{x}} - \max(Z_{\mathbf{x}_1^\star}, \ldots, Z_{\mathbf{x}_r^\star})\right] \geq \kappa\sqrt{\frac{L_k}{2^\alpha (2r)^{\alpha/d} \log(r)}},$$

for some universal constant $\kappa > 0$. back

More about noisy kriging-based optimization

D. Huang, T. T. Allen, W.I. Notz, and N. Zeng (2006).

Global optimization of stochastic black-box systems via sequential kriging meta-models.

Journal of Global Optimization, 34:441-466.

W. Scott, P. Frazier, and W. Powell (2011).

The correlated knowledge gradient for simulation optimization of continuous parameters using gaussian process regression.

SIAM Journal on Optimization, 21:996-1026.

V. Picheny, D. Ginsbourger, Y. Richet and G. Caplin (2013).

Quantile-based optimization of noisy computer experiments with tunable precision.

Technometrics, Volume 55(1), pp. 2-36 (with discussion and rejoinder).

V. Picheny and D. Ginsbourger (2014)

Noisy kriging-based optimization methods: A unified implementation within the DiceOptim package.

Computational Statistics and Data Analysis, Volume 71, pp.1035-1053.

back