Bayesian excursion set estimation with GPs

Dario Azzimonti [1]

Joint works with: David Ginsbourger, Clément Chevalier, Julien Bect, Yann Richet

[1]Institute of Mathematical Statistics and Actuarial Science
University of Bern, Switzerland (Visiting UQOD group, Idiap Research Institute)

Uncertainty Quantification Workshop
Sheffield, UK

September 15, 2016

Outline

Outline

The framework

In this talk we focus on the problem of determining the set

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

where $D \subset \mathbb{R}^d$ is compact, $f : D \longrightarrow \mathbb{R}^k$ is measurable, $T \subset \mathbb{R}^k$.

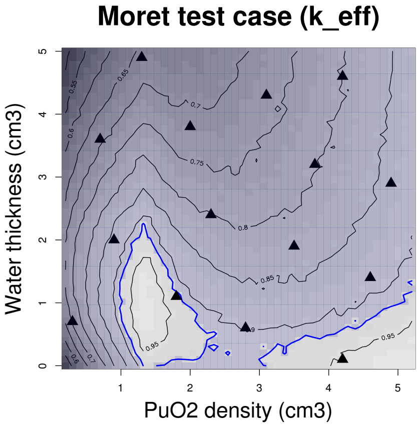
Here: $k = 1$, f is continuous, and $T = (-\infty, t]$ for a fixed $t \in \mathbb{R}$.

$\Gamma^\star = \{x \in D : f(x) \leq t\}$ is denoted the excursion set of f below t.

Objective

Estimate $\Gamma^\star$ and quantify uncertainty on it when f is evaluated only at a few points $\mathbf{X}_n = \{x_1, \ldots, x_n\} \subset D$.

The framework: IRSN test case



Test case:

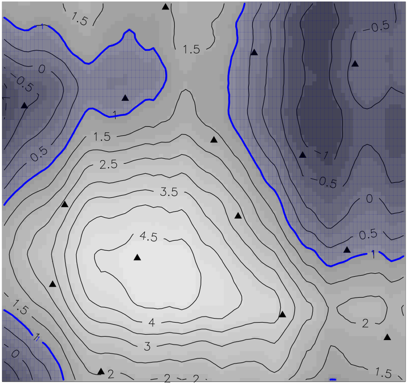
- $\mathrm{k_{eff}}$ function of $\mathrm{PuO_2}$ density and $\mathrm{H_2O}$ thickness, $D = [0.2, 5.2] \times [0, 5]$;
- continuous function, expensive to evaluate;
- $n = 20$ observations (black triangles);

Objective: estimate $\Gamma^\star = \{x \in D : f(x) \leq t\}$ and evaluate the uncertainty of the estimate.

Acknowledgements: Yann Richet, Institut de Radioprotection et de Sûreté Nucleaire.

The framework: an example

Gaussian random field realization



Function $f : D \subset \mathbb{R}^d \to \mathbb{R}$

- expensive to evaluate;
- continuous.

Evaluated at
$\mathbf{X}_n = (x_1, \dots, x_n)$ (black triangles)
with values $\mathbf{f}_n = (f(x_1), \dots, f(x_n))$.

Objective: estimate
$\Gamma^\star = \{x \in D : f(x) \leq t\}$ and
evaluate the uncertainty of the estimate.

Bayesian approach

Bayesian framework: f *is seen as one realization of a (GRF)* $(Z_x)_{x \in D}$ with prior mean m and covariance kernel k.

Given the function evaluations $\mathbf{f}_n$ the posterior field has a Gaussian distribution

$$Z \mid (Z(\mathbf{X}_n) = \mathbf{f}_n)$$

with mean and covariance kernel

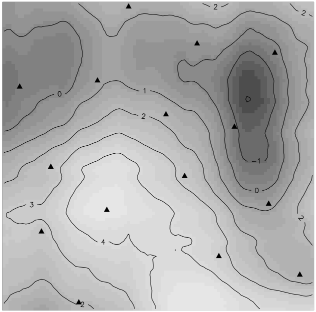
$$m_n(x) = m(x) + k(x, \mathbf{X}_n) k(\mathbf{X}_n, \mathbf{X}_n)^{-1} (\mathbf{f}_n - m(\mathbf{X}_n))$$
$$k_n(x, y) = k(x, y) - k(x, \mathbf{X}_n) k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, y)$$

$\Gamma^\star$ is a realization of $\Gamma = \{x \in D : Z_x \leq t\} = Z^{-1}((-\infty, t])$

A prior on the space of functions

Assume: f realization of $(Z_x)_{x \in D}$, Gaussian Random Field (GRF)

Posterior GRF realization



Prior: $(Z_x)_{x \in D}$ with

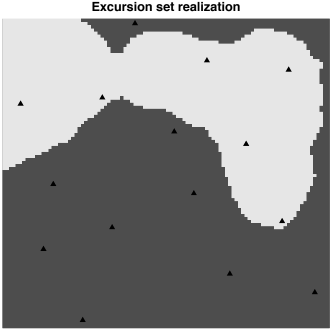
- a.s. continuous paths;
- Matérn covariance kernel k $(\nu = 3/2)$;
- constant mean function m.

Given $n = 15$ evaluations $\mathbf{f}_n$ at $\mathbf{X}_n$

Posterior field: $Z \mid Z_{\mathbf{X}_n} = \mathbf{f}_n$
with mean m_n and covariance k_n.

Distribution of excursion sets

The posterior field defines posterior distribution on excursion sets.
$\Gamma = \{x \in D : Z_x \leq t\}$



How to summarize the distribution on sets?

The posterior excursion set is a random closed set.

Here we focus on **Expectations** of random closed sets[1]

- Vorob'ev expectation
- distance average expectation

Conservative estimates, based on Vorob'ev quantiles.

1. for more definitions of expectation see Molchanov, I. (2005). Theory of Random Sets. Springer.

Main references:

E. Vazquez and M. P. Martinez. (2006). *Estimation of the volume of an excursion set of a Gaussian process using intrinsic kriging.* Tech Report. arXiv:math/0611273.

Ranjan, P., Bingham, D., and Michailidis, G. (2008). *Sequential experiment design for contour estimation from complex computer codes.* Technometrics, 50(4):527541.

Bect, J., Ginsbourger, D., Li, L., Picheny, V., and Vazquez, E. (2012). *Sequential design of computer experiments for the estimation of a probability of failure.* Stat. Comput., 22 (3):773793.

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

Chevalier, C., Ginsbourger, D., Bect, J., and Molchanov, I. (2013). *Estimating and quantifying uncertainties on level sets using the Vorobev expectation and deviation with Gaussian process models.* mODa 10.

Bolin, D. and Lindgren, F. (2015), French, J. P. and Sain, S. R. (2013)

and references therein...

Outline

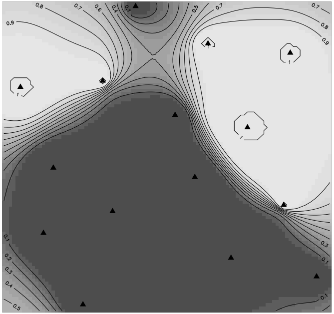
Vorob'ev quantiles

The function

$$p_n : x \in D \to p_n(x) = P_n(x \in \Gamma) \in [0, 1]$$

is the **coverage function of** Γ, where $P_n(\cdot) = P(\cdot \mid Z_{\mathbf{X}_n} = \mathbf{f}_n)$.

Coverage probability function

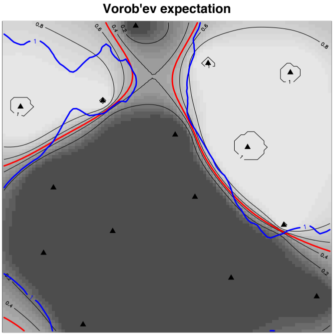


In the Gaussian case

- fast to compute
  $p_n(x) = \Phi\left(\frac{m_n(x) - t}{\sqrt{k_n(x,x)}}\right)$
- marginal statement
- creates a family of set estimates
  $Q_\rho = \{x \in D : p_n(x) \geq \rho\}$

Vorob'ev expectation

Consider a Borel measure μ on D. From the family of "quantiles" Q_ρ we can choose $Q_{\widetilde{\rho}}$ such that $\mu(Q_{\widetilde{\rho}}) = \mathbb{E}[\mu(\Gamma)]$.



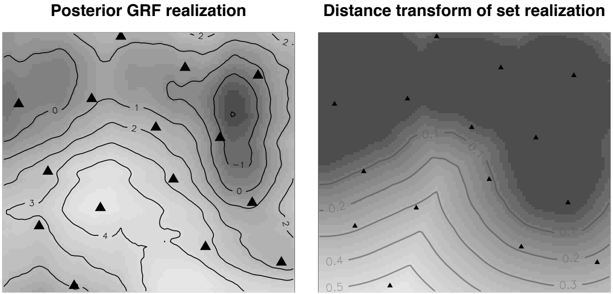
Properties:

- based on the measure μ;
- for some choices of μ, fast to compute;
- no confidence statements on the set.

Chevalier, C., Ginsbourger, D., Bect, J., and Molchanov, I. (2013). *Estimating and quantifying uncertainties on level sets using the Vorobev expectation and deviation with Gaussian process models.* mODa 10

Distance average approach

Consider the distance function $d : (x, \Gamma) \to d(x, \Gamma)$.



Γ is random therefore $d(x, \Gamma)$ is a random variable for each $x \in D$.

Distance average expectation

Given the distance function $d(x, \Gamma)$, the expected distance function

$$\overline{d}(x) = \mathbb{E}[d(x, \Gamma)]$$

The distance average expectation of Γ is the set

$$\mathbb{E}_{DF}[\Gamma] = \{x \in D : \overline{d}(x) \leq \overline{\varepsilon}\} \quad \text{where}$$

$\overline{\varepsilon}$ is chosen in order to obtain a distance function for the set $\mathbb{E}_{DF}[\Gamma]$ as "close" as possible to $\overline{d}$ in a L^2 sense.

An **uncertainty assessment** for the estimate is

$$\mathrm{DFV}_{\Gamma} = \mathbb{E}\|\overline{d}(\cdot) - d(\cdot, \Gamma)\|_2^2$$

Distance average expectation

Given the distance function $d(x, \Gamma)$, the expected distance function

$$\overline{d}(x) = \mathbb{E}[d(x, \Gamma)]$$

The distance average expectation of Γ is the set

$$\mathbb{E}_{DF}[\Gamma] = \{x \in D : \overline{d}(x) \leq \overline{\varepsilon}\} \quad \text{where}$$

$\overline{\varepsilon}$ is chosen in order to obtain a distance function for the set $\mathbb{E}_{DF}[\Gamma]$ as "close" as possible to $\overline{d}$ in a L^2 sense.

An **uncertainty assessment** for the estimate is

$$\mathrm{DFV}_{\Gamma} = \mathbb{E}\|\overline{d}(\cdot) - d(\cdot, \Gamma)\|_2^2$$

Distance average expectation

Given the distance function $d(x, \Gamma)$, the expected distance function

$$\overline{d}(x) = \mathbb{E}[d(x, \Gamma)]$$

The distance average expectation of Γ is the set

$$\mathbb{E}_{DF}[\Gamma] = \{x \in D : \overline{d}(x) \leq \overline{\varepsilon}\} \quad \text{where}$$

$\overline{\varepsilon}$ is chosen in order to obtain a distance function for the set $\mathbb{E}_{DF}[\Gamma]$ as "close" as possible to $\overline{d}$ in a L^2 sense.

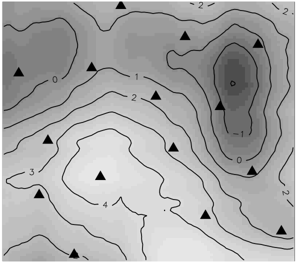
An **uncertainty assessment** for the estimate is

$$\mathrm{DFV}_{\Gamma} = \mathbb{E}\|\overline{d}(\cdot) - d(\cdot, \Gamma)\|_2^2$$

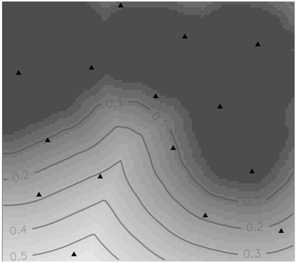
Distance average approach

Consider the distance function $d : (x, \Gamma) \to d(x, \Gamma)$.

Posterior GRF realization

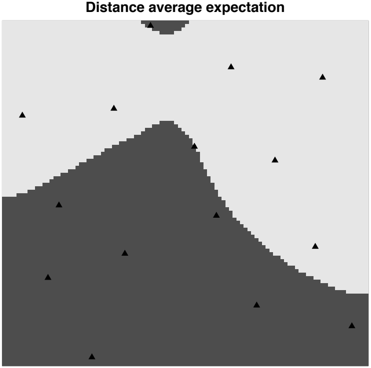


Distance transform of set realization



For each realization of Γ (expensive!) we compute the distance function and then consider an average over the functions.

Distance average expectation



$$\mathbb{E}_{DF}[\mathit{\Gamma}] = \{x \in D : \overline{d}(x) \leq \overline{\varepsilon}\}$$

Outline

An approximate Gaussian random field

Assumption: the GRF Z has been evaluated at $x_1, \ldots, x_n \in D$.

We denote by $Z_E = (Z_{e_1}, \ldots, Z_{e_m})'$ the random vector of values of Z at $E = \{e_1, \ldots, e_m\} \subset D$.

Here we focus on affine predictors of Z of the form

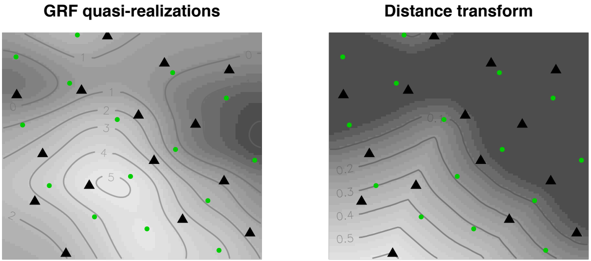
$$\widetilde{Z}_x = a(x) + \mathbf{b}^T(x) Z_E \quad (x \in D),$$

where $a : D \longrightarrow \mathbb{R}$ is a trend function and $\mathbf{b} : D \longrightarrow \mathbb{R}^m$ is a vector-valued function of deterministic weights.

Similarly, we approximate Γ by the excursion set of $\widetilde{Z}$:
$\widetilde{\Gamma} = \{x \in D : \widetilde{Z}_x \leq t\}$

Towards an optimal design of simulation points

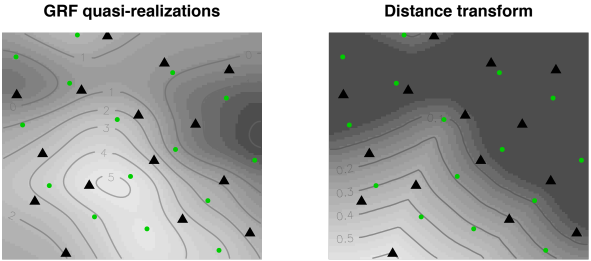
The simulation points E could be chosen with a LHS design ($m = 30$)



However, we do not control on how close is Γ to $\widetilde{\Gamma}$

Towards an optimal design of simulation points

The simulation points E could be chosen with a LHS design ($m = 30$)



However, we do not control on how close is Γ to $\widetilde{\Gamma}$

What distance between Γ to $\widetilde{\Gamma}$?

Definition: the function

$$(\Gamma_1, \Gamma_2) \in D \times D \longrightarrow d_{\mu,n}(\Gamma_1, \Gamma_2) = \mathbb{E}[\mu(\Gamma_1 \Delta \Gamma_2) \mid Z_{\mathbf{x}_n} = \mathbf{f}_n]$$

is called expected distance in measure between Γ_1, Γ_2.

Proposition: distance in measure between Γ and $\widetilde{\Gamma}$

a) If Z and $\widetilde{Z}$ are random fields such that Γ and $\widetilde{\Gamma}$ are random closed sets, $D \subset \mathbb{R}^d$ is compact and μ is a finite Borel measure on D, we have

$$d_{\mu,n}(\Gamma, \widetilde{\Gamma}) = \int \rho_{n,m}(x)\mu(dx) \text{ where}$$

$$\rho_{n,m}(x) = P_n(x \in \Gamma \Delta \widetilde{\Gamma}) = \mathbb{P}_n(Z_x \geq t, \widetilde{Z}_x < t) + \mathbb{P}_n(Z_x < t, \widetilde{Z}_x \geq t)$$

Towards an optimal design of simulation points

Proposition: distance in measure between Γ and $\widetilde{\Gamma}$

b) If Z is Gaussian conditionally on $Z_{\mathbf{X}_n}$ with conditional mean $\mathfrak{m}_n$ and conditional covariance kernel k_n, we get

$$\mathbb{P}_n(Z_x \geq t, \widetilde{Z}_x < t) = \Phi_2\left(\mathbf{c}_n(x, E), \Sigma_n(x, E)\right),$$

with $\mathbf{c}_n(x, E) = \begin{pmatrix} \mathfrak{m}_n(x) - t \\ t - a(x) - \mathbf{b}(x)^T \mathfrak{m}_n(E) \end{pmatrix}$

and $\Sigma_n(x, E) = \begin{pmatrix} k_n(x, x) & -\mathbf{b}(x)^T k_n(E, x) \\ -\mathbf{b}(x)^T k_n(E, x) & \mathbf{b}(x)^T k_n(E, E)\mathbf{b}(x) \end{pmatrix}$

Towards an optimal design of simulation points

c) Particular case: If $\widetilde{Z}$ is chosen as best linear unbiased predictor of Z given $Z(\mathbf{X}_n)$, then $\mathbf{b}(x) = k_n(E,E)^{-1}k_n(E,x)$ so that

$$\Sigma_n(x,E) = \begin{pmatrix} k_n(x,x) & -\gamma_n(x,E) \\ -\gamma_n(x,E) & \gamma_n(x,E) \end{pmatrix}$$

where $\gamma_n(x,E) = \mathrm{Var}_n[\widehat{Z}_x] = k_n(E,x)^T k_n(E,E)^{-1} k_n(E,x)$.

Optimal design(s) of simulation points can be obtained by minimizing

$$d_{\mu,n}(\Gamma,\widetilde{\Gamma}(E)) = \int \Phi_2\left(\mathbf{c}_n(x,E),\Sigma_n(x,E)\right) + \Phi_2\left(-\mathbf{c}_n(x,E),\Sigma_n(x,E)\right)\mu(\mathrm{d}x)$$

over $(\mathbf{e}_1,\ldots,\mathbf{e}_m) \in D^m$.

Towards an optimal design of simulation points

c) Particular case: If $\widetilde{Z}$ is chosen as best linear unbiased predictor of Z given $Z(\mathbf{X}_n)$, then $\mathbf{b}(x) = k_n(E,E)^{-1}k_n(E,x)$ so that

$$\Sigma_n(x,E) = \begin{pmatrix} k_n(x,x) & -\gamma_n(x,E) \\ -\gamma_n(x,E) & \gamma_n(x,E) \end{pmatrix}$$

where $\gamma_n(x,E) = \mathrm{Var}_n[\widehat{Z}_x] = k_n(E,x)^T k_n(E,E)^{-1} k_n(E,x)$.

Optimal design(s) of simulation points can be obtained by minimizing

$$d_{\mu,n}(\Gamma, \widetilde{\Gamma}(E)) = \int \Phi_2\left(\mathbf{c}_n(x,E), \Sigma_n(x,E)\right) + \Phi_2\left(-\mathbf{c}_n(x,E), \Sigma_n(x,E)\right) \mu(\mathrm{d}x)$$

over $(\mathbf{e}_1, \ldots, \mathbf{e}_m) \in D^m$.

Procedure overview

Approximate Z at each point with

$$\widetilde{Z}_x = a(x) + \mathbf{b}^T(x) Z_E \text{ with } E = \{\mathbf{e}_1, \dots, \mathbf{e}_m\}$$

The points in E are chosen with one of the following algorithms:

Algorithm A (Full criterion): sequential minimization of

$$d_{\mu,n}(\Gamma, \widetilde{\Gamma}(E_i^*)) = \int \Phi_2 \left(\mathbf{c}_n(x, E_i^*), \Sigma_n(x, E_i^*)\right) + \Phi_2 \left(-\mathbf{c}_n(x, E_i^*), \Sigma_n(x, E_i^*)\right) \mu(\mathrm{d}x)$$

with respect to $\mathbf{e}_i$ where $E_i^* = \{\mathbf{e}_1^*, \dots, \mathbf{e}_{i-1}^*\} \cup \{\mathbf{e}_i\}$;

Algorithm B (Fast heuristic): sequential maximization of

$$\rho_{n,E}(x) = \Phi_2 \left(\mathbf{c}_n(x, E), \Sigma_n(x, E)\right) + \Phi_2 \left(-\mathbf{c}_n(x, E), \Sigma_n(x, E)\right)$$

with respect to x;

Procedure overview

Approximate Z at each point with

$$\widetilde{Z}_x = a(x) + \mathbf{b}^T(x) Z_E \text{ with } E = \{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$$

The points in E are chosen with one of the following algorithms:

Algorithm A (Full criterion): sequential minimization of

$$d_{\mu,n}(\Gamma, \widetilde{\Gamma}(E_i^*)) = \int \Phi_2\left(\mathbf{c}_n(x, E_i^*), \Sigma_n(x, E_i^*)\right) + \Phi_2\left(-\mathbf{c}_n(x, E_i^*), \Sigma_n(x, E_i^*)\right) \mu(\mathrm{d}x)$$

with respect to $\mathbf{e}_i$ where $E_i^* = \{\mathbf{e}_1^*, \ldots, \mathbf{e}_{i-1}^*\} \cup \{\mathbf{e}_i\}$;

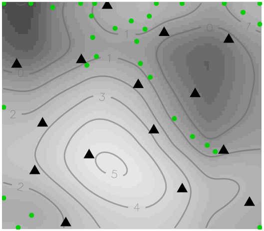
Algorithm B (Fast heuristic): sequential maximization of

$$\rho_{n,E}(x) = \Phi_2\left(\mathbf{c}_n(x, E), \Sigma_n(x, E)\right) + \Phi_2\left(-\mathbf{c}_n(x, E), \Sigma_n(x, E)\right)$$

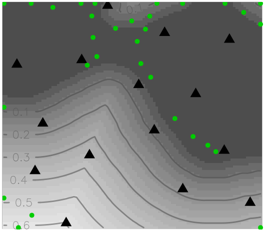
with respect to x;

Quasi realizations for distance average variability

GRF quasi-realizations



Distance transform



A.D. and Bect, J. and Chevalier, C. and Ginsbourger, D. (2016) *Quantifying uncertainties on excursion sets under a Gaussian random field prior*. SIAM/ASA J.Uncertainty Quantification, 4(1):850–874. hal-01103644v2.

Test case: negative Branin-Hoo function

Quantity of interest: $\mathrm{DTV}_{\Gamma,N} = \frac{1}{N}\sum_{i=1}^{N}\|d_N^*(\cdot) - d(\cdot,\Gamma_i)\|_2^2$

Experimental set-up:

- 20 observation points;
- $N = 10000$ conditional simulations on a 50×50 grid;
- $K = 100$ replications of each experiment.

Methods:

1. Full Monte Carlo simulations on the grid,
2. Simulations at optimized points (A,B) and interpolation on the same grid.

Test case: negative Branin-Hoo function

Method 1: Full grid simulations

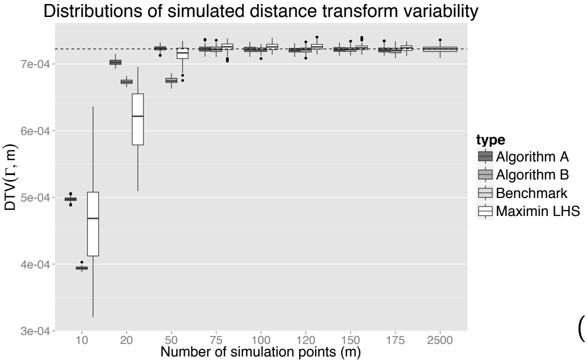
- Variability:

$$\widehat{\mathbb{E}}\left[\mathrm{DTV}_{\Gamma,10000}\right] = 7.2 \times 10^{-4} (\pm 4.71 \times 10^{-8});$$
$$\widehat{\mathrm{Var}}\left[\mathrm{DTV}_{\Gamma,10000}\right] = 2.22 \times 10^{-11} (\pm 3.14 \times 10^{-13});$$

- total computational cost: **10498 seconds**.

Test case: negative Branin-Hoo function

Method 2: quasi-realizations on 50×50 grid.



Total computing cost

- Algorithm A, $m = 150$: **11201 sec** (10566 for simulation point optimization)
- Algorithm B, $m = 150$: **812 sec** (250 for simulation point optimization)
- LHS, $m = 150$: **691 sec**

(Benchmark: **10498 sec**)

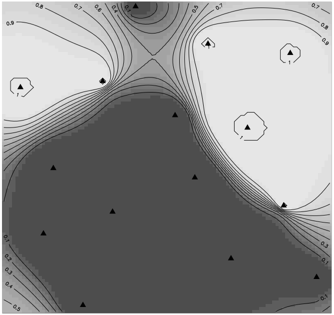
Outline

Vorob'ev quantiles

The coverage function $p_n : x \to p_n(x) = P_n(x \in \Gamma)$ defines the family of set estimates

$$Q_\rho = \{x \in D : p_n(x) \geq \rho\}$$

Coverage probability function

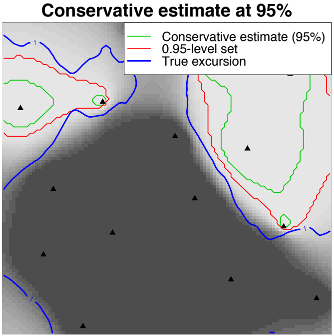


If $\rho = \widetilde{\rho}$ we have Vorob'ev expectation

High values of ρ gives us sets with high marginal probability of observing the set.

A **conservative estimate** of Γ^* is

$$C_{\Gamma,n} = Q_{\rho^*} \text{ where } \rho^* \in \arg\max_{\rho\in[0,1]}\{\mu(Q_\rho) : P_n(Q_\rho \subset \{Z_x \leq t\}) \geq \alpha\}$$



- joint confidence statement on the set estimate;
- method introduced for Gauss Markov random fields;
- expensive to compute otherwise.

Bolin, D. and Lindgren, F. (2015). *Excursion and contour uncertainty regions for latent Gaussian models.* JRSS: B, 77(1):85-106.

The computation of conservative estimates

The family of sets Q_ρ is *nested*, therefore we can obtain $C_{\Gamma,n}$ with a **dichotomy on the level** ρ.

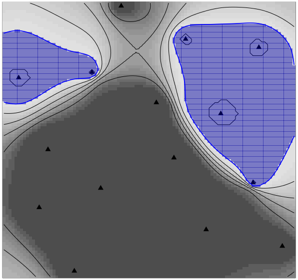
At each iteration of the dichotomy we need to compute

$$P_n(Q_\rho \subset \{Z_x \leq t\}) = P_n(Z_{e_1} \leq t, \dots, Z_{e_k} \leq t),$$

where $E = \{e_1, \dots, e_k\}$ is a the discretization of Q_ρ.

- randomized quasi Monte Carlo integration by Genz et al. :
  (Fast, reliable, dimension dependent, available only $k < 1000$)
- standard Monte Carlo.
  (dimension independent, many samples for low variance)

A quasi Monte Carlo algorithm for orthant probabilities



$$P_n(Q_\rho \subset \{Z_x \leq t\}) =$$
$$P_n(Z_{e_1} \leq t, \ldots, Z_{e_k} \leq t) =$$
$$1 - P_n(\max_{x \in E} Z_x > t) = 1 - p$$

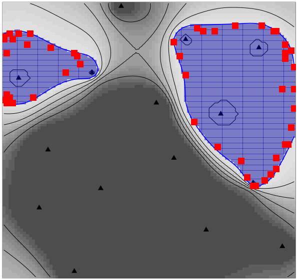
Main idea: $p = P_n(\max_E Z_x > t) = p_q + (1 - p_q) R_q$, where

$$p_q = P_n(\max_{E_q} Z_x > t), \qquad R_q = P_n(\max_{E \setminus E_q} Z_x > t \mid \max_{E_q} Z_x \leq t).$$

Genz algorithm (QRSVN) — Monte Carlo methods

A quasi Monte Carlo algorithm for orthant probabilities

Candidate $\rho = 0.95$



$$P_n(Q_\rho \subset \{Z_x \leq t\}) =$$
$$P_n(Z_{e_1} \leq t, \dots, Z_{e_k} \leq t) =$$
$$1 - P_n(\max_{x \in E} Z_x > t) = 1 - p$$

Main idea: $p = P_n(\max_E Z_x > t) = p_q + (1 - p_q) R_q$, where

$$p_q = P_n(\max_{E_q} Z_x > t),$$
Genz algorithm (QRSVN)

$$R_q = P_n(\max_{E \setminus E_q} Z_x > t \mid \max_{E_q} Z_x \leq t).$$
Monte Carlo methods

$$\widehat{p_q} = 0.47 \qquad \widehat{R_q} = 0.42 \qquad \Rightarrow \widehat{p} = 0.69$$

Computation of the remainder

$R_q = P_n(\max_{E \setminus E_q} Z_x > t \mid \max_{E_q} Z_x \leq t)$

Standard Monte Carlo:

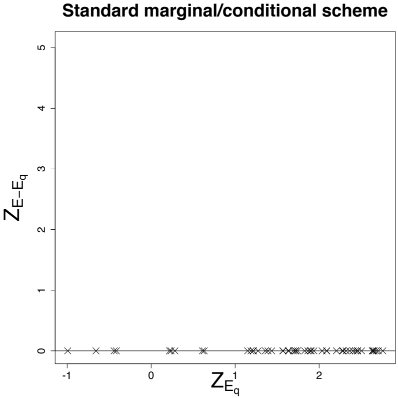
1. draw realizations $z_1^q, \ldots, z_s^q$ from $Z_{E_q} \mid \max_{E_q} Z_x \leq t$;
2. for each z_i^q, draw a realization from $Z_{E \setminus E_q} \mid Z_{E_q} = z_i^q$;
3. Estimate R_q with $R_q^{\text{MC}} = \frac{1}{s} \sum_{i=1}^{s} \mathbf{1}_{\max(Z_{E \setminus E_q}(\omega_i) \mid Z_{E_q} = z_i^q) > t}$

The cost of step 1 is higher than the cost of step 2.

At fixed computational budget we reduce the variance of R_q^{MC} exploiting this difference with **asymmetric nested Monte Carlo**.

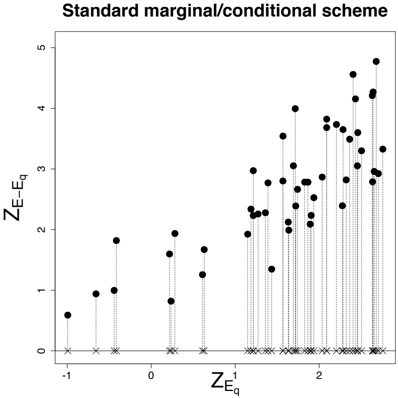
Computation of the remainder

At fixed computational budget we reduce the variance of R_q^{MC} drawing many realizations of $Z_{E \setminus E_q} \mid Z_{E_q} = z_i^q$ for each z_i.



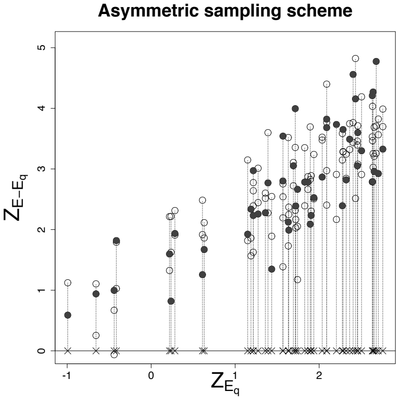
Computation of the remainder

At fixed computational budget we reduce the variance of R_q^{MC} drawing many realizations of $Z_{E \setminus E_q} \mid Z_{E_q} = z_i^q$ for each z_i.



Computation of the remainder

At fixed computational budget we reduce the variance of R_q^{MC} drawing many realizations of $Z_{E \setminus E_q} \mid Z_{E_q} = z_i^q$ for each z_i.



Computation of the remainder: asymmetric nested MC

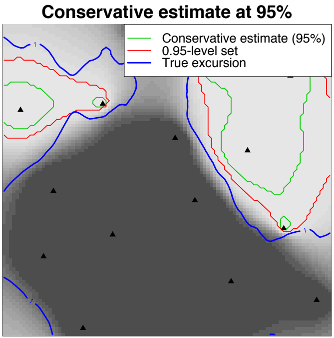
$R_q = P_n(\max_{E \setminus E_q} Z_x > t \mid \max_{E_q} Z_x \leq t)$

1. *draw realizations* $z_1^q, \ldots, z_s^q$ *from* $Z_{E_q} \mid \max_{E_q} Z_x \leq t$;
2. *for each* z_i^q, **draw** $m^* > 1$ **samples** from $Z_{E \setminus E_q} \mid Z_{E_q} = z_i^q$;
3. $R_q^{\mathrm{anMC}} = \frac{1}{s} \frac{1}{m^*} \sum_{i=1}^{s} \sum_{j=1}^{m^*} \mathbf{1}_{\max(Z_{E \setminus E_q}(\omega_{i,j}) \mid Z_{E_q} = z_i^q) > t}$

$\mathrm{var}(R_q^{\mathbf{anMC}})$ **is optimally reduced if:** $m^* = \sqrt{\frac{(\alpha+c)B}{\beta(A-B)}}$,
where $A = \mathrm{var}(\mathbf{1}_{\max(Z_{E \setminus E_q} \mid Z_{E_q}) > t})$, $B = \mathbb{E}\left[\mathrm{var}(\mathbf{1}_{\max(Z_{E \setminus E_q} \mid Z_{E_q}) > t} \mid \max_{E_q} Z_x \leq t)\right]$ and α, β, c system dependent constants.

A D. and Ginsbourger D. (2016). *Estimating orthant probabilities of high dimensional Gaussian vectors with an application to set estimation*. Submitted, hal-01289126

Comparison with standard Monte Carlo

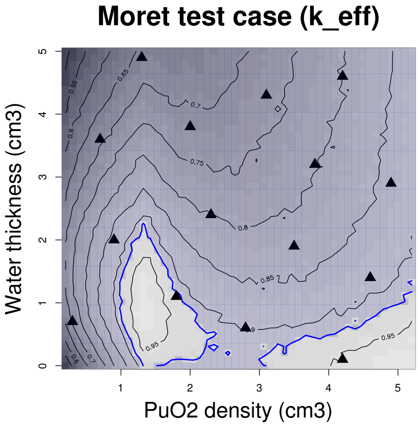


Full discretization: grid 100×100

Time for equivalent estimates:

- Full MC: 1520 seconds;
- GMC: 200 seconds;
- GanMC: 136 seconds.

The IRSN test case



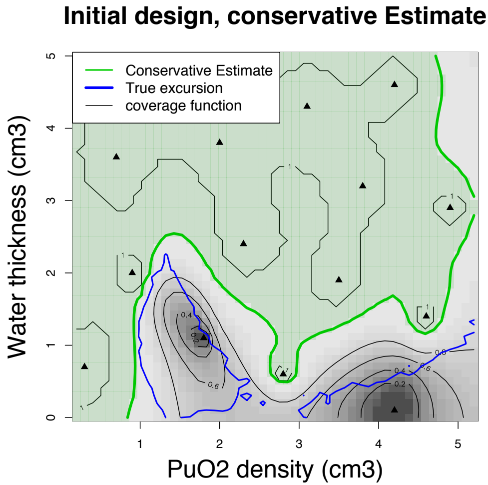
Test case:

- $\mathrm{k}_{\mathrm{eff}}$ function of PuO_2 density and $\mathrm{H}_2\mathrm{O}$ thickness, $D = [0.2, 5.2] \times [0, 5]$;
- $n = 20$ observations (LHS design);
- $\Gamma^* = \{x \in D : \mathrm{k}_{\mathrm{eff}}(x) \leq t\}$, $t = 0.92$

GRF model

- constant prior mean, Matérn ($\nu = 5/2$) covariance;
- MLE for parameters.

Acknowledgements: Yann Richet, Institut de Radioprotection et de Sûreté Nucleaire.



Discretization on grid 50×50.

Conservative estimate at 95%;

Candidate sets dimension between 1659 and 2084;

Volume of conservative estimate: 17.36 (true volume 22.0). Sequential

Conclusion

- GP can be used for uncertainty quantification on sets;
- different types of estimates, depending on the final objective;
- Optimal quasi-realizations for excursion sets lower the computational cost of quantities based on set realizations;
- Conservative estimates:
  - sequential strategies to reduce uncertainty;
  - GanMC: benchmark study with other algorithms;
  - Currently developing R package ConservativeEstimates.

Thanks for your attention!

Conclusion

- GP can be used for uncertainty quantification on sets;
- different types of estimates, depending on the final objective;
- Optimal quasi-realizations for excursion sets lower the computational cost of quantities based on set realizations;
- Conservative estimates:
  - sequential strategies to reduce uncertainty;
  - GanMC: benchmark study with other algorithms;
  - Currently developing R package ConservativeEstimates.

Thanks for your attention!

References

A D. and Ginsbourger D. (2016). *Estimating orthant probabilities of high dimensional Gaussian vectors with an application to set estimation*. Submitted, hal-01289126

A D. and Bect, J. and Chevalier, C. and Ginsbourger, D. (2015). *Quantifying uncertainties on excursion sets under a Gaussian random field prior*. Accepted, JUQ hal-01103644v2

Bolin, D. and Lindgren, F. (2015). *Excursion and contour uncertainty regions for latent Gaussian models.* JRSS: B, 77(1):85-106.

Chevalier, C., Ginsbourger, D., Bect, J., and Molchanov, I. (2013). *Estimating and quantifying uncertainties on level sets using the Vorobev expectation and deviation with Gaussian process models.* mODa 10.

How to reduce the uncertainty on the estimate?

Stepwise uncertainty reduction: find a sequence of evaluation points $X_1, X_2, \dots$ that optimally reduces the expected uncertainty on the future estimate, i.e. given an initial design $\mathbf{X}_n$, select

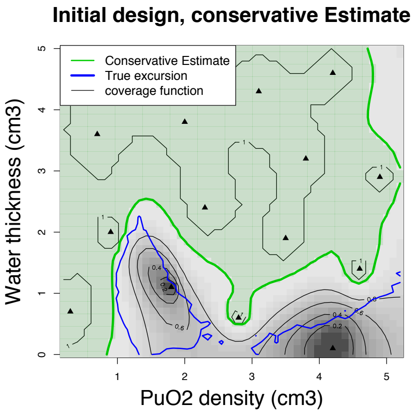
$$X_{n+1} \in \arg\min_{x_{n+1} \in D} \mathbb{E}_n[H_{n+1} \mid X_{n+1} = x_{n+1}]$$

Uncertainty function(s): many possible definitions, here

$$H_{n+1}^{\text{symm}} = \mathbb{E}_{n+1}[\mu(\Gamma \Delta Q_{\rho_{n+1}})], \quad \Gamma \Delta Q_{\rho_{n+1}} = \Gamma \setminus Q_{\rho_{n+1}} \cup Q_{\rho_{n+1}} \setminus \Gamma$$

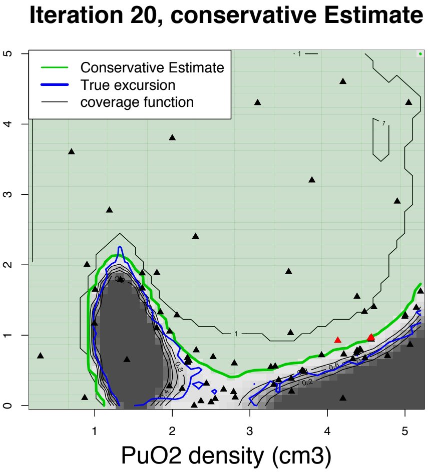
see [Bect et al. (2012), Chevalier et al. (2014)] and references therein for different definitions of H_{n+1}.

How to reduce the uncertainty on the estimate?



Criterion: $J_{n+q}^{\text{symm}}(\mathbf{x}_q) = \mathbb{E}_n[\mathbb{E}_{n+q}[\mu(\Gamma \setminus C_{\Gamma,n})] \mid \mathbf{X}_{n+q} = \mathbf{x}_q],$

Sequential strategies



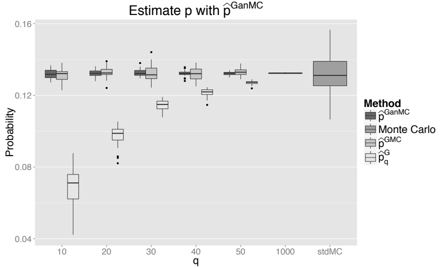
$n = 75$ new evaluations;

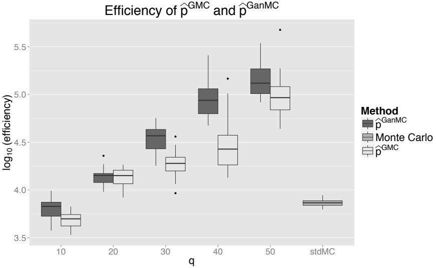
next evaluation chosen in order to minimize the future expected volume of the set difference $\Gamma \setminus Q_{\rho^*}$;

Volume of updated CE: 20.72 (true excursion: 22.0, old estimate: 17.36) Back

Joint work with: David Ginsbourger, Clément Chevalier, Julien Bect, Yann Richet.

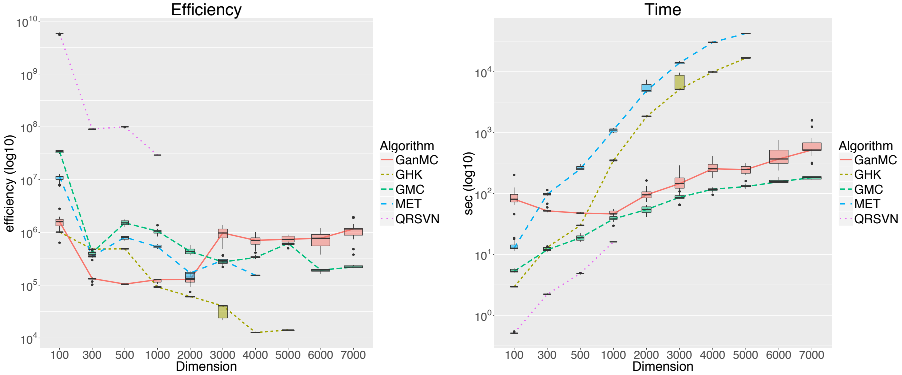
More comparisons anMC/MC





Benchmark: 6d GRF, discretization: 1000 Sobol' points, k Matérn ($\nu = 5/2$) with $\theta = [0.5, 0.5, 1, 1, 0.5, 0.5]^T$ and $\sigma^2 = 8$, m constant.

More comparisons anMC/MC



Benchmark: 6d GRF, discretization: 1000 Sobol' points, k Matérn ($\nu = 5/2$) with $\theta = [0.5, 0.5, 1, 1, 0.5, 0.5]^T$ and $\sigma^2 = 8$, m constant, $t = 5$.