Quantization-based LHS for dependent inputs

Application to BUVARD-MES and sensitivity analysis.

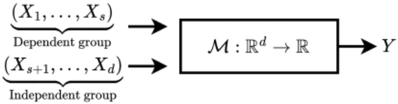
Guerlain Lambert
Céline Helbert[1] & Claire Lauvernet[2]

[1] Univ. Lyon, UMR CNRS 5208 École Centrale de Lyon ICJ, [2]INRAE, RIVERLY, Villeurbanne

25 Septembre 2024

Context and Motivation

Consider a function $\mathcal{M} : \mathbb{R}^d \to \mathbb{R}$ such that :



Objective : We aim to estimate $\mathbb{E}[\mathcal{M}(X)]$.

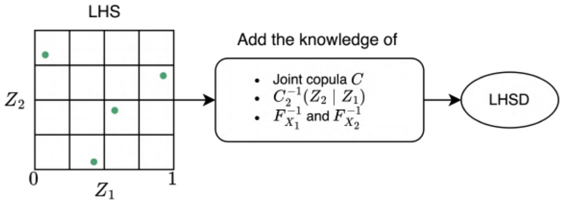
Existing methods :

- Monte Carlo sampling
- LHS with rank correlation [Iman and Conover, 1980, Stein, 1987]
- Kernel Herding [Chen et al., 2012]

1. Context and motivation
2. Design of computer experiment
   - Existing method
   - Quantization-based LHS
3. Numerical experiment

Latin Hypercube Sampling with Dependence (LHSD)

Existing sampling design based on **LHS** [McKay et al., 1979] to take into account the dependence of the joint distribution via copulas [Mondal and Mandal, 2020].



Issue : Requires knowledge of a copula, conditional copula (and its inverse) and the i.c.d.f. of the marginals.

Propose a sampling methodology with minimal prerequisites to account for dependence.

Vector quantification in a nutshell

Quantization ? Approach a random vector X by a N-quantizer $\Gamma = \{x_1, \ldots, x_N\} \in \left(\mathbb{R}^d\right)^N$:

$$\hat{X} = \mathrm{Proj}(X, \Gamma) = \sum_{i=1}^{N} x_i 1_{\{X \in C_i(\Gamma)\}} \text{ with } C_i \text{ : Voronoï partition}$$

Let $\hat{X}$ associated to the N-quantizer $\Gamma = \{x_1, \ldots, x_N\}$, $\mathbb{N} \in \mathbb{N}^*$.

- We call **distortion** at level N, the mapping from $\left(\mathbb{R}^d\right)^N \to \mathbb{R}$:

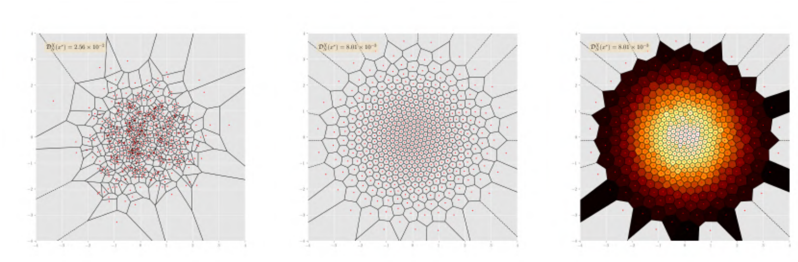
$$\mathcal{D}_N^X(\Gamma) = \|X - \hat{X}\|_2^2 = \mathbb{E}\left[\min_{1 \le i \le N} |X - x_i|^2\right]$$

- Any N-quantizer satisfying the following minimization problem is said to be **optimal** :

$$\inf_{\Gamma, \#\Gamma \le N} \mathcal{D}_N^X(\Gamma) = \inf\left\{\|X - \hat{X}\|_2^2, \Gamma \subset \mathbb{R}^d, \#\Gamma \le N\right\}$$

An example of an optimal quantification application

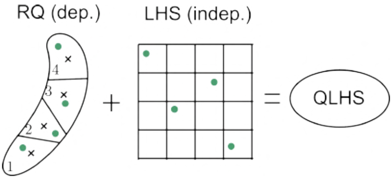
- **Context :** Application to the normal distribution $\mathcal{N}(0, I_2)$
- **Optimal quantization :** difficult to obtain in practice
- Use of a stochastic algorithm : Competitive Learning Vector Quantization (**CLVQ**)



(a) Initial (b) Sans Poids (c) Avec Poids

Ref : [Pagès, 2015]

Quantization-based LHS



- $\Gamma^*_{X_{dep}}$: optimal quantization of X_{dep}
- $U_i \sim \mathcal{L}(X_{dep} \mid X_{dep} \in C_i(\Gamma^*_{X_{dep}}))$
- V_i : LHS sample for X_{indep}

- Unbiased estimator based on QLHS :

$$\mu_{QLHS} := \sum_{i=1}^{N} \mathbb{P}[X_{dep} \in C_i(\Gamma^*_{X_{dep}})] f(U_i, V_{\pi(i)})$$

Reference : [Lambert et al., 2024]

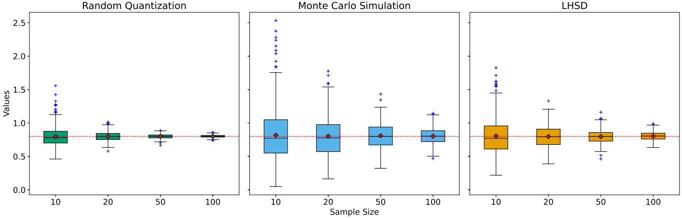
Without LHS, for $\delta > 0$, there exists a constant $\tilde{C}_{d,\delta} > 0$ such that :

$$\inf_{(C_i) \in \mathcal{C}_I^N} \sup_{L_f \leq 1} \mathrm{Var}(\mu_{RQ}) \leq \tilde{C}_{d,\delta} N^{-2/d} \|X\|_2^2$$

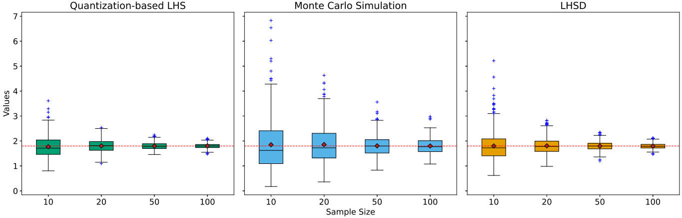
Based on [Corlay and Pagès, 2010]

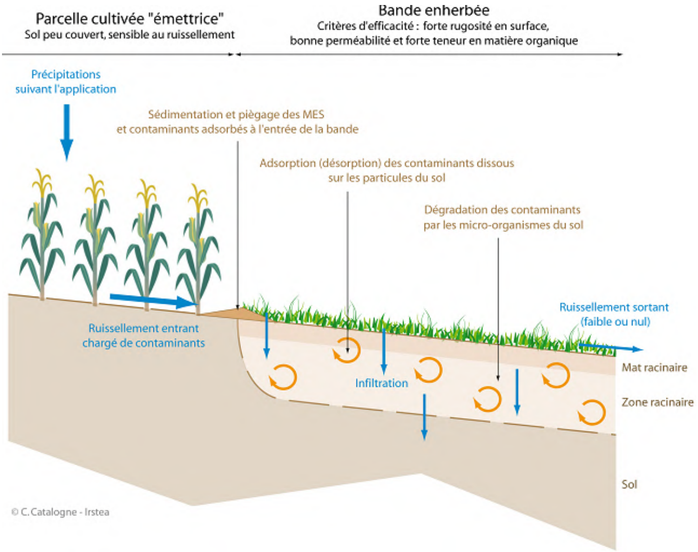
1. Context and motivation
2. Design of computer experiment
3. Numerical experiment
   - Toy functions
   - A design tool for vegetated strips : BUVARD_MES
   - Water retention curve estimation
   - HSIC test on the dependent group

Random Quantization : $\mathbb{E}[X_1 X_2]$ where $X = (X_1, X_2)$ is a centred Gaussian vector with $cov(X_1, X_2) = 0.8$.



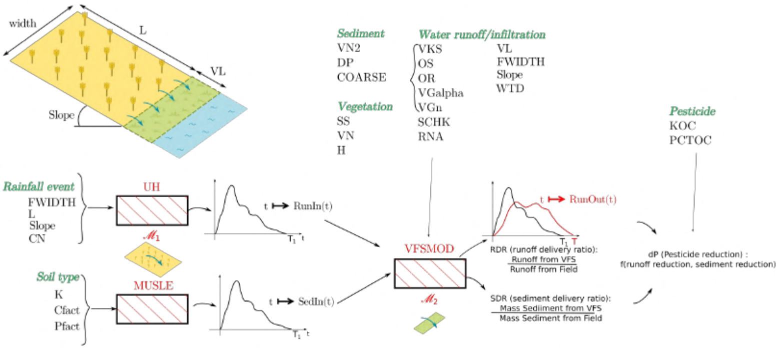
QLHS : $\mathbb{E}[(X_1 + X_2)^2 X_{indep}]$ where $X_{dep} = (X_1, X_2)$ is a centred Gaussian vector with $cov(X_1, X_2) = 0.8$ and $X_{indep} \sim \mathcal{U}([0, 1])$.





Reference : [Veillon et al., 2022]

Schematic diagram of BUVARD_MES

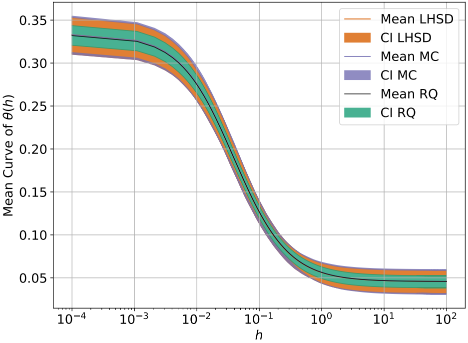


Reference : [Catalogne et al., 2018, Veillon et al., 2022]

Water retention curve estimation

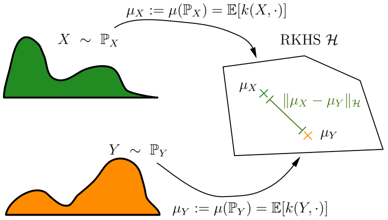
Van Genuchten' Water retention curve :

$$\theta(h) = \theta_r + \frac{\theta_s - \theta_r}{\left(1 + (\alpha|h|)^n\right)^{1-1/n}}$$



Mean Water retention curve estimation by MC, LHSD, RQ with $N = 10$.

Application to the BUVARD_MES model



- Estimation of HSIC via V-stat
- Independence test between : $\mathrm{HSIC}(X_i, Y) = 0 \Leftrightarrow X_i \perp Y$

Input	p-value				Decision			
	Ref	MC	QLHS	LHSD	Ref	MC	QLHS	LHSD
Van Genuchten	0	0.18	0.031	0.096	✓	✗	✓	✗

Sensitivity Analysis (screening HSIC). Settings : Ref ($N = 10000$, asymp. test), {MC, QLHS, LHSD} ($N = 400$, perm. test)

Results : QLHS performs better than LHSD (LHS+copula) and crude Monte Carlo.

Conclusion and outlook

What has been done

1. QLHS provides an effective sampling strategy for dependent inputs.
2. The benefits of quantization compared with an LHSD plan.

Outlooks

- Derive an explicit variance reduction factor for QLHS.
- Penalized the distortion loss function to obtain equiprobable Voronoï cells.

Catalogne, C., Lauvernet, C., and Carluer, N. (2018).
Guide d'utilisation de l'outil buvard pour le dimensionnement des bandes tampons végétalisées destinées à limiter les transferts de pesticides par ruissellement.

Chen, Y., Welling, M., and Smola, A. (2012).
Super-samples from kernel herding.

Corlay, S. and Pagès, G. (2010).
Functional quantization-based stratified sampling methods.
Monte Carlo Methods and Applications, 21:1 – 32.

Iman, R. and Conover, W. (1980).
Small sample sensitivity analysis techniques for computer models.with an application to risk assessment.
9(17):1749–1842.
Publisher: Taylor & Francis.

Lambert, G., Helbert, C., and Lauvernet, C. (2024).
Quantization-based LHS for dependent inputs : application to sensitivity analysis of environmental models.
working paper or preprint.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979).
A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code.
Technometrics, 21(2):239–245.
Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

Mondal, A. and Mandal, A. (2020).
Stratified random sampling for dependent inputs in monte carlo simulations from computer experiments.
Journal of Statistical Planning and Inference, 205:269–282.

Pagès, G. (2015).
Introduction to vector quantization and its applications for numerics.
ESAIM: Proceedings and Surveys, 48:29–79.

Stein, M. (1987).
Large sample properties of simulations using latin hypercube sampling.
29(2):143–151.

Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

Veillon, F., Carluer, N., Lauvernet, C., and Rabotin, M. (2022).
BUVARD-MES : Un outil en ligne pour dimensionner les zones tampons enherbées afin de limiter les transferts de pesticides vers les eaux de surface.
In *50e congrès du Groupe Français des Pesticides*.