GSA of Model Output With Dependent Input: Part I Sampling Techniques

Thierry A. Mara

University of Reunion

12^{th} SAMO Summer School
June 24-28 2024
University of Parma

Outline

Notations

Introduction

Case 1: Marginal and conditional cdfs known

The Rosenblatt transformation

A special case: the Nataf transformation

Case 2: Marginal and conditional pdfs unknown

Rejection sampling (MCMC)

Notations & Assumptions

- $y = f(\boldsymbol{x})$ the scalar QoI
- $\boldsymbol{x} = (x_1, x_2, \dots, x_d) \sim p_x = \dfrac{\partial^d F_x}{\partial x_1 \partial \dots \partial x_d}$
- $x_i \sim p_{x_i} = \dfrac{\mathrm{d}F_{x_i}}{\mathrm{d}x_i}$, the marginal pdf ($F_{x_i}$ cdf) of x_i
- $p_{x_i|x_j} = \dfrac{\mathrm{d}F_{x_i|x_j}}{\mathrm{d}x_i}$, the conditional pdf (cdf) of x_i over x_j, ...
- $\mathcal{D} = \{0, 1, \dots, d\}$, $\mathcal{D}_{-i} \bigcap \{i\} = \emptyset$, $\mathcal{D}_{+i} \bigcap \{i\} \neq \emptyset$
- $\boldsymbol{x} = (\boldsymbol{x}_\alpha, \boldsymbol{x}_{-\alpha})$ with $\boldsymbol{x}_\alpha \bigcap \boldsymbol{x}_{-\alpha} = \emptyset$
- $\alpha = \{i_1, \dots, i_k\} \subseteq \mathcal{D} \Leftrightarrow \boldsymbol{x}_\alpha = \boldsymbol{x}_{i_1,\dots,i_k} = (x_{i_1}, \dots, x_{i_k})$

Assumptions: $\mathbb{E}\left[f^2\right] = \int_{\mathbb{R}^d} f^2(\boldsymbol{x}) p_x \mathrm{d}\boldsymbol{x} < \infty$

Introduction

When $\boldsymbol{x} = (x_1, \ldots, x_d) \sim p_{x_1} p_{x_2} \ldots p_{x_d}$, we saw that it was always possible to turn the problem into the following: $y = g(\boldsymbol{u})$ with $\boldsymbol{u} = (u_1, \ldots, u_d) \sim \mathcal{U}\,(0,1)^d$ by setting $u_i = F_{x_i}(x_i)$.

What-if $p_x = p_{x_1} p_{x_2} \ldots p_{x_d}$?

Case 1: Marginal and conditional cdfs known

RT

<u>The Rosenblatt Transform</u>: Suppose known all cdfs:
$(F_{x_{i_1}}, F_{x_{i_2}|x_{i_1}}, F_{x_{i_3}|x_{i_1},x_{i_2}}, \ldots, F_{x_{i_d}|\boldsymbol{x}_{\sim i_d}})$, $\forall i_k \in (1, 2, \ldots, d)$.

RT

<u>The Rosenblatt Transform</u>: Suppose known all cdfs: $(F_{x_{i_1}}, F_{x_{i_2}|x_{i_1}}, F_{x_{i_3}|x_{i_1},x_{i_2}}, \ldots, F_{x_{i_d}|\boldsymbol{x}_{\sim i_d}})$, $\forall i_k \in (1, 2, \ldots, d)$. Then one can use the Rosenblatt transform (1952) to sample $\boldsymbol{x}$ from $\boldsymbol{u}$ (or the other way around),

$$\begin{cases} x_{i_1} & = & F^{-1}_{x_{i_1}}(u_{i_1}) \\ x_{i_2} & = & F^{-1}_{x_{i_2}}(u_{i_2}|u_{i_1}) \\ \vdots & \vdots & \vdots \\ x_{i_d} & = & F^{-1}_{x_{i_d}|\boldsymbol{x}_{\sim i_d}}(u_{i_d}|\boldsymbol{u}_{\sim i_d}) \end{cases} \tag{1}$$

<u>N.B.</u>: **The Rosenblatt transformation is not unique** as there are $d!$ possible transformations.

Example

We want a sample of $(x_1, x_2) \in \mathcal{U}(0,1)^2$ uniformly distributed over the triangle $x_1 + x_2 \leq 1$. The problem being symmetric, we have $F_{x_1} = F_{x_2}$ and $F_{x_1|x_2} = F_{x_2|x_1}$.

Example

We want a sample of $(x_1, x_2) \in \mathcal{U}(0,1)^2$ uniformly distributed over the triangle $x_1 + x_2 \leq 1$. The problem being symmetric, we have $F_{x_1} = F_{x_2}$ and $F_{x_1|x_2} = F_{x_2|x_1}$.

It can be shown that the marginal cdf is:
$F_{x_i}(x_i) = 1 - (1 - x_i)^2 = u_i$

and the conditional cdf is: $F_{x_j|x_i}(x_i, x_j) = \frac{x_j}{1-x_i} = u_j$

Example

We want a sample of $(x_1, x_2) \in \mathcal{U}(0,1)^2$ uniformly distributed over the triangle $x_1 + x_2 \leq 1$. The problem being symmetric, we have $F_{x_1} = F_{x_2}$ and $F_{x_1|x_2} = F_{x_2|x_1}$.

It can be shown that the marginal cdf is:
$F_{x_i}(x_i) = 1 - (1 - x_i)^2 = u_i$

and the conditional cdf is: $F_{x_j|x_i}(x_i, x_j) = \frac{x_j}{1-x_i} = u_j$

To sample $\boldsymbol{x}$ from a sample of $\boldsymbol{u}$ we must make the following transformations,

$$\begin{cases} x_i &= 1 - \sqrt{1 - u_i} \\ x_j &= u_j\sqrt{1 - u_i} \end{cases} \qquad (i,j) = (1,2) \text{ or } (i,j) = (2,1)$$

Cholesky transformation

$$\text{From } \boldsymbol{u} \sim \mathcal{U}(0,1)^d$$

$$\text{to}$$

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

NT

Let $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a random vector of RVs normally distributed. $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ is the vector of means and $\boldsymbol{\Sigma}$ is a $d \times d$ Covariance Matrix (symmetric & positive-definite). If $\boldsymbol{\Sigma}$ is diagonal, then the RVs are **independent** otherwise they are **correlated**.

NT

Let $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a random vector of RVs normally distributed. $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ is the vector of means and $\boldsymbol{\Sigma}$ is a $d \times d$ Covariance Matrix (symmetric & positive-definite). If $\boldsymbol{\Sigma}$ is diagonal, then the RVs are **independent** otherwise they are **correlated**.

We note that

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^T} \tag{2}$$

with $|.|$ is the determinant.

NT

Let $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a random vector of RVs normally distributed. $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ is the vector of means and $\boldsymbol{\Sigma}$ is a $d \times d$ Covariance Matrix (symmetric & positive-definite). If $\boldsymbol{\Sigma}$ is diagonal, then the RVs are **independent** otherwise they are **correlated**.

We note that

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^T} \tag{2}$$

with $|.|$ is the determinant.

By setting $\boldsymbol{z} = (\boldsymbol{x} - \boldsymbol{\mu})\boldsymbol{U}^{-1}$ where $\boldsymbol{U}$ is the **upper triangular Cholesky matrix** defined as $\boldsymbol{\Sigma} = \boldsymbol{U}^T\boldsymbol{U}$, Eq.(2) becomes,

$$\mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I}_d) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T} \tag{3}$$

which means that $\boldsymbol{z}$ is a vector of **independent standard normal variables**.

NT

Let $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a random vector of RVs normally distributed. $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ is the vector of means and $\boldsymbol{\Sigma}$ is a $d \times d$ Covariance Matrix (symmetric & positive-definite). If $\boldsymbol{\Sigma}$ is diagonal, then the RVs are **independent** otherwise they are **correlated**.

We note that

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^T} \tag{2}$$

with $|.|$ is the determinant.

By setting $\boldsymbol{z} = (\boldsymbol{x} - \boldsymbol{\mu})\boldsymbol{U}^{-1}$ where $\boldsymbol{U}$ is the **upper triangular Cholesky matrix** defined as $\boldsymbol{\Sigma} = \boldsymbol{U}^T\boldsymbol{U}$, Eq.(2) becomes,

$$\mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I}_d) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\boldsymbol{z}\boldsymbol{z}^T} \tag{3}$$

which means that $\boldsymbol{z}$ is a vector of **independent standard normal variables**. We get $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{z}\boldsymbol{U} \Rightarrow$ Samples of $\boldsymbol{x}$ can be generated from samples of $\boldsymbol{z}$ which can be generated from $\boldsymbol{u}$.

Cholesky transformation

Exercises

Exercise 1: We want a sample of $\mathcal{N}\left(\boldsymbol{x}\middle|\begin{bmatrix}-1\\1\end{bmatrix},\begin{bmatrix}2 & 1\\1 & 2\end{bmatrix}\right)$. Set $N = 128$, and generate a sample of $(u_1, u_2) \sim \mathcal{U}(0,1)^2$.

1. Transform u_1 into $z_1 \sim \mathcal{N}(z_1|0,1)$
2. Transform u_2 into $z_2 \sim \mathcal{N}(z_2|0,1)$
3. Find $\boldsymbol{U}$ the upper Cholesky matrix of the covariance matrix
4. Deduce a sample of $\boldsymbol{x}$. Check the empirical covariance matrix.

Nataf transformation

From $\boldsymbol{u} \sim \mathcal{U}(0,1)^d$ to

$$\boldsymbol{x} \sim c.p_{x_1}.p_{x_2}.p_{x_3}\dots p_{x_d}$$

with c a Gaussian copula density

Nataf transformation

The Nataf Transform: Let $\boldsymbol{x} = (x_1, \ldots, x_d)$ be a random vector of correlated RVs distributed w.r.t. $(F_{x_1}, \ldots, F_{x_d})$ with correlation matrix $\boldsymbol{C}_{xx}$.

Nataf transformation

<u>The Nataf Transform</u>: Let $\boldsymbol{x} = (x_1, \ldots, x_d)$ be a random vector of correlated RVs distributed w.r.t. $(F_{x_1}, \ldots, F_{x_d})$ with correlation matrix $\boldsymbol{C}_{xx}$. To generate samples of $\boldsymbol{x}$, Iman & Conover (1982) have proposed the following algorithm

1. Let $\boldsymbol{u} \sim \mathcal{U}(0,1)^d$
2. Transform $\boldsymbol{u}$ into $\tilde{\boldsymbol{x}}$ with the integral transform method, that is, $\tilde{x}_i = F_{x_i}^{-1}(u_i)$
3. Transform $\boldsymbol{u}$ into $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{C}_{xx})$ with the Cholesky transformation
4. Transform $\tilde{\boldsymbol{x}}$ into $\boldsymbol{x}$ such that $\mathrm{rank}(x_i) = \mathrm{rank}(z_i)$, $\forall i = 1, \ldots, d$

<u>N.B.</u>: Iman & Conover's method ensures that $\boldsymbol{x}$ and $\boldsymbol{z}$ has the same Rank (Spearman) Correlation Matrix. If the target is the (Pearson) Correlation Matrix, then one might need to modify $\boldsymbol{C}_{xx}$ in Step 3. In that case, the technique is known as the Nataf Transform (1962).

About the Rank Transformation:

Let consider the following samples,

$$\tilde{x}_1 = \begin{bmatrix} 0.55 \\ 0.31 \\ 0.78 \\ 0.03 \\ 0.27 \end{bmatrix}, \; z_1 = \begin{bmatrix} -0.15 \\ 0.61 \\ 0.38 \\ -0.31 \\ 0.91 \end{bmatrix}$$

Their ranks are resp.,

$$\mathrm{rank}(\tilde{x}_1) = \begin{bmatrix} 4 \\ 3 \\ 5 \\ 1 \\ 2 \end{bmatrix}, \; \mathrm{rank}(z_1) = \begin{bmatrix} 2 \\ 4 \\ 3 \\ 1 \\ 5 \end{bmatrix}$$

About the Rank Transformation:

Let consider the following samples,

$$\tilde{x}_1 = \begin{bmatrix} 0.55 \\ 0.31 \\ 0.78 \\ 0.03 \\ 0.27 \end{bmatrix}, \quad z_1 = \begin{bmatrix} -0.15 \\ 0.61 \\ 0.38 \\ -0.31 \\ 0.91 \end{bmatrix}$$

Their ranks are resp.,

$$\mathrm{rank}(\tilde{x}_1) = \begin{bmatrix} 4 \\ 3 \\ 5 \\ 1 \\ 2 \end{bmatrix}, \quad \mathrm{rank}(z_1) = \begin{bmatrix} 2 \\ 4 \\ 3 \\ 1 \\ 5 \end{bmatrix}$$

They would have the same rank, for instance, by rearranging $\tilde{x}_1$ as follows,

$$x_1 = \begin{bmatrix} 0.27 \\ 0.55 \\ 0.31 \\ 0.03 \\ 0.78 \end{bmatrix} \to \mathrm{rank}(x_1) = \begin{bmatrix} 2 \\ 4 \\ 3 \\ 1 \\ 5 \end{bmatrix}$$

Exercises

Exercise: We want a sample of (x_1, x_2) with $p_{x_1} = \mathcal{U}(1,2)$, $p_{x_2} = \mathcal{N}(0,2)$ and (Pearson) Correlation Matrix $\boldsymbol{C}_{xx} = \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}$. Set $N = 128$, and generate a sample of $(u_1, u_2) \sim \mathcal{U}(0,1)^2$. Set $\boldsymbol{R}_{xx} = \boldsymbol{C}_{xx}$

1. Transform the sample of $\boldsymbol{u}$ into a sample of $\tilde{\boldsymbol{x}} \sim p_{x_1}.p_{x_2}$
2. Transform the sample of $\boldsymbol{u}$ into a sample of $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{R}_{xx})$
3. Rank-transform $\tilde{\boldsymbol{x}}$ to obtain $\boldsymbol{x}$
4. Check the (Pearson) correlation matrix of $\boldsymbol{x}$. If not satisfactory change $\boldsymbol{R}_{xx}$ and go to Step 2.

With Rosenblatt we perform the following transformation: $\boldsymbol{u} \rightarrow \boldsymbol{x}$
With Nataf transformation we do: $\boldsymbol{u} \rightarrow \boldsymbol{z} \rightarrow \boldsymbol{x}$
In the end of the day we can write:

$$y = f(\boldsymbol{x}) = g(\boldsymbol{u}) = h(\boldsymbol{z})$$

With Rosenblatt we perform the following transformation: $\boldsymbol{u} \to \boldsymbol{x}$
With Nataf transformation we do: $\boldsymbol{u} \to \boldsymbol{z} \to \boldsymbol{x}$
In the end of the day we can write:

$$y = f(\boldsymbol{x}) = g(\boldsymbol{u}) = h(\boldsymbol{z})$$

With the pick-freeze method (seen on Monday morning) to compute the Sobol' indices how would you proceed:

$$S_i^{IA} = \frac{2\sum_{n=1}^{N}\left(y_n^A - y_n^{A_{b_i}}\right)\left(y_n^{B_{a_i}} - y_n^B\right)}{\sum_{n=1}^{N}\left(y_n^A - y_n^B\right)^2 + \left(y_n^{A_{b_i}} - y_n^{B_{a_i}}\right)^2}$$

$$T_i^{IA} = \frac{\sum_{n=1}^{N}\left(y_n^A - y_n^{A_{b_i}}\right)^2 + \left(y_n^{B_{a_i}} - y_n^B\right)^2}{\sum_{n=1}^{N}\left(y_n^A - y_n^B\right)^2 + \left(y_n^{A_{b_i}} - y_n^{B_{a_i}}\right)^2}$$

With Rosenblatt we perform the following transformation: $\boldsymbol{u} \rightarrow \boldsymbol{x}$
With Nataf transformation we do: $\boldsymbol{u} \rightarrow \boldsymbol{z} \rightarrow \boldsymbol{x}$
In the end of the day we can write:

$$y = f(\boldsymbol{x}) = g(\boldsymbol{u}) = h(\boldsymbol{z})$$

How would you generate the output samples: $(\boldsymbol{y}^A, \boldsymbol{y}^B, \boldsymbol{y}^{A_{b_i}}, \boldsymbol{y}^{B_{a_i}})$?

With Rosenblatt we perform the following transformation: $\boldsymbol{u} \rightarrow \boldsymbol{x}$
With Nataf transformation we do: $\boldsymbol{u} \rightarrow \boldsymbol{z} \rightarrow \boldsymbol{x}$
In the end of the day we can write:

$$y = f(\boldsymbol{x}) = g(\boldsymbol{u}) = h(\boldsymbol{z})$$

Answer: Generate $(\boldsymbol{u}^A, \boldsymbol{u}^B, \boldsymbol{u}^{A_{b_i}}, \boldsymbol{u}^{B_{a_i}})$, transform them into $(\boldsymbol{x}^A, \boldsymbol{x}^B, \boldsymbol{x}^{A_{b_i}}, \boldsymbol{x}^{B_{a_i}})$, run the model and collect the model responses which correspond to $(\boldsymbol{y}^A, \boldsymbol{y}^B, \boldsymbol{y}^{A_{b_i}}, \boldsymbol{y}^{B_{a_i}})$.

With Rosenblatt we perform the following transformation: $\boldsymbol{u} \rightarrow \boldsymbol{x}$
With Nataf transformation we do: $\boldsymbol{u} \rightarrow \boldsymbol{z} \rightarrow \boldsymbol{x}$
In the end of the day we can write:

$$y = f(\boldsymbol{x}) = g(\boldsymbol{u}) = h(\boldsymbol{z})$$

But how do we interpret the sensitivity indices of the u-variables?
Besides, RT is not unique, which RT shall we use?
Answer in Part II.

Cholesky transformation

Compute the variance-based sensitivity indices of the linear function:

$$y = \sum_{i=1}^{3} x_i$$

with $p(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$
where $\boldsymbol{\mu} = (1, 2, 3)$ and

$$\boldsymbol{C} = \begin{array}{cc} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} & \\ \begin{bmatrix} 1 & -0.5 & 0 \\ -0.5 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{bmatrix} & \begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} \end{array} \quad \boldsymbol{C} = \begin{array}{cc} \begin{array}{ccc} x_2 & x_3 & x_1 \end{array} & \\ \begin{bmatrix} 1 & 0.8 & -0.5 \\ 0.8 & 1 & 0 \\ -0.5 & 0 & 1 \end{bmatrix} & \begin{array}{c} x_2 \\ x_3 \\ x_1 \end{array} \end{array}$$

Generate the samples of $\boldsymbol{u}$ with $N = 128$. Compute the Sobol' indices (S_i^{IA}, T_i^{IA}) with the RT (here Cholesky transformation) of (u_1, u_2, u_3) into (x_1, x_2, x_3). What-if we RT transform (u_2, u_3, u_1) into (x_2, x_3, x_1)?

Case 2: Marginal and conditional cdfs unknown

Rejection Sampling: Let $\boldsymbol{x}$ be a random vector of RVs distributed w.r.t. the joint pdf $p_{\boldsymbol{x}}$. If none of the techniques above can be applied: Try acceptance/rejection sampling techniques like Markov Chains Monte Carlo (MCMC).

But, ignoring the independent u-variables, the sensitivity analysis that can be performed is limited.

Some References

- M.D. McKay, R.J. Beckman, and W.J. Conover (1979). Technometrics, 239–245.
- Nataf, B. (1962), Compt. Rend. Acad. Sci., 42–43
- Iman, R.I., Conover, W.J. (1982). Commun. Stat. Simul. Comput. 311–334.
- Rosenblatt, M (1952). Annals Math. Stat., 470–472