New Perspectives for Sensitivity Analysis

Sebastien Da Veiga – Safran Tech

Journées GDR 06/30/2016
OUTLINE

- Context

- Generalized GSA
  - Distances between probability distributions
  - RKHS embedding
  - Orthogonal decompositions

- Conclusion & Perspectives
Sensitivity Analysis

Goal: identify and rank the input parameters according to their impact on the output of a computer code

Why?
- Reduce the output uncertainty efficiently by reducing the uncertainty of the main contributors
- Improve the knowledge of the physical phenomenon,
- Simplify the model

Notations

\[ Y = \eta(X_1, \ldots, X_d) \]

Output
Input parameters
Two points of view

- **Local Sensitivity**: studies the behavior of the output locally around a nominal value of the inputs

  \[
  S_i = \frac{\sigma^2_i}{\text{Var}(Y)} \left( \frac{\partial \eta(X)}{\partial X_i} \bigg|_{X=X_0} \right)^2
  \]

  - Easy to compute and apprehend
  - But local approach, turns global only if the model is linear

- **Global sensitivity**: all input parameters vary in their uncertain domain and we analyze the output variations

There are links between the viewpoints when local sensitivity is repeated (DGSM, Lamboni et al. 2013)
Global Sensivity Analysis (GSA) – 2 families of methods

- Screening methods
  - Standard DOEs
  - Sequential bifurcation, ...
  - Morris

\[ n \approx d/2 - 10d \]

- Quantitative methods based on a variance decomposition
  - Linear regression, SRC, ...
  - Sobol indices

\[ n \approx 2d - 10^4 d \]
CONTEXT

GSA – Focus on Sobol indices
- Sobol-Hoeffding decomposition for independent input parameters

\[
\eta(X) = \eta_0 + \sum_{i=1}^{d} \eta_i(X_i) + \sum_{1 \leq i < j \leq d} \eta_{i,j}(X_i, X_j) + \ldots + \eta_1,\ldots,d(X_1, \ldots, X_d)
\]

- Functions are centered and orthogonal
- Formulas with conditional expectations:

\[
\eta_0 = \mathbb{E}(Y)
\]
\[
\eta_i(X_i) = \mathbb{E}(Y|X_i) - \mathbb{E}(Y)
\]
\[
\eta_{i,j}(X_i, X_j) = \mathbb{E}(Y|X_i, X_j) - \mathbb{E}(Y|X_i) - \mathbb{E}(Y|X_j) + \mathbb{E}(Y)
\]
\ldots
**CONTEXT**

**GSA – Focus on Sobol indices**
- By orthogonality

\[
\text{Var}(\eta(X)) = \sum_{i=1}^{d} \text{Var}(\eta_i(X_i)) + \sum_{1 \leq i < j \leq d} \text{Var}(\eta_i,j(X_i, X_j)) + \ldots + \text{Var}(\eta_1,\ldots,d(X_1, \ldots, X_d))
\]

- The total variance is decomposed into pieces involving main effects, 2\textsuperscript{nd} order interactions, and so on
- \( \Rightarrow \) Possibility to define the sensitivity index of a group of input parameters

\[
S_I(X_I) = \frac{\text{Var}(\eta_I(X_I))}{\text{Var}(\eta(X))}
\]

\[
S_i(X_i) = \frac{\text{Var}(\mathbb{E}(Y|X_i))}{\text{Var}(Y)} \quad \text{Main effect}
\]

\[
S_i^T(X_i) = \sum_{I \supseteq i} S_I \quad \text{Total effect}
\]
Limitations

- Variance decomposition is just a particular (and limited) analysis of the output variation

- The numerical code is expensive to evaluate
  - Usually rely on surrogate model to estimate Sobol indices

- The number of input parameters may be large (100 – 1000)
  - In practice, a first screening step is necessary

- Inputs & outputs may not be scalars (curves, …)
 CONTEXT

 ➔ Limitations

- Variance decomposition is just a particular (and limited) analysis of the output variation

- The numerical code is expensive to evaluate
  - Usually rely on surrogate model to estimate Sobol indices

- The number of input parameters may be large (100 – 1000)
  - In practice, a first screening step is necessary

- Inputs & outputs may not be scalars (curves, …)

Take Home Message I
Generalized GSA
Limitations

- Variance decomposition is just a particular (and limited) analysis of the output variation

- The numerical code is expensive to evaluate
  - Usually rely on surrogate model to estimate Sobol indices

- The number of input parameters may be large (100 – 1000)
  - In practice, a first screening step is necessary

- Inputs & outputs may not be scalars (curves, …)

Take Home Message I
Generalized GSA

Take Home Message II
Links between generalized GSA and feature selection …
Limitations

- Variance decomposition is just a particular (and limited) analysis of the output variation
- The numerical code is expensive to evaluate
  - Usually rely on surrogate model to estimate Sobol indices
- The number of input parameters may be large (100 – 1000)
  - In practice, a first screening step is necessary
- Inputs & outputs may not be scalars (curves, …)

Take Home Message I
Generalized GSA

Take Home Message II
Links between generalized GSA and feature selection …

... which can accommodate structured objects
OUTLINE

→ Context

→ **Generalized GSA**
  - Distances between probability distributions
  - RKHS embedding
  - Orthogonal decompositions

→ Conclusion & Perspectives
GENERALIZED GSA

Going beyond the variance decomposition

- « Jitter » the input probability distributions (Lemaître et al. 2015)
- Indices based on contrast functions (Fort et al. 2014)

\[
S_i^\psi = \mathbb{E}_i \psi(Y; \theta^*) - \mathbb{E}_{(X_i, Y)} \psi(Y; \theta_i(X_i))
\]

\[
\theta^* = \arg \min_\theta \mathbb{E} \psi(Y; \theta)
\]

\[
\theta_i(x) = \arg \min_\theta \mathbb{E} (\psi(Y; \theta) | X_i = x)
\]
GENERALIZED GSA

- Going beyond the variance decomposition
  - « Jitter » the input probability distributions (Lemaître et al. 2015)
  - Indices based on contrast functions (Fort et al. 2014)

\[ S^\psi_i = \mathbb{E}_\psi(Y; \theta^*) - \mathbb{E}_{(X_i, Y)} \psi(Y; \theta_i(X_i)) \]
\[ \theta^* = \arg \min_\theta \mathbb{E}_\psi(Y; \theta) \]
\[ \theta_i(x) = \arg \min_\theta \mathbb{E}(\psi(Y; \theta)|X_i = x) \]

- Quantify the impact of an input parameter on the probability distribution of the output

\[ S^{TV}_i = \int |p_Y(y) - p_{Y|X_i=x}(y)| p_{X_i}(x) dx dy \]  
Borgonovo 2007

\[ S^{KL}_i = \int p_{Y|X_i=x}(y) \ln \left( \frac{p_{Y|X_i=x}(y)}{p_Y(y)} \right) p_{X_i}(x) dx dy \]  
Kraskov et al. 2001
GENERALIZED GSA

General framework for distributional indices

- From a broad perspective, the impact of an input parameter may be defined through the choice of a similarity measure between probability distributions

$$S_i = \mathbb{E}_{X_i} \left(d(P_Y, P_{Y|X_i})\right)$$

- If the input probability distribution and the conditional one are « close », the input parameter has little influence
GENERALIZED GSA

- General framework for distributional indices

- From a broad perspective, the impact of an input parameter may be defined through the choice of a similarity measure between probability distributions

\[ S_i = \mathbb{E}_{X_i} \left( d(P_Y, P_{Y|X_i}) \right) \]

- If the input probability distribution and the conditional one are « close », the input parameter has little influence

- Toy example

\[ Y = \sin(X_1) + 5\sin^2(X_2) + 0.1X_3^4\sin(X_1) \]

\[ X_1, X_2, X_3, X_4 \sim U(-\pi, \pi) \]

Ishigami function with dummy variable
GENERALIZED GSA
GENERALIZED GSA
What do you think?
GENERALIZED GSA

How can we compare probability distributions?

- The basics
GENERALIZED GSA

How can we compare probability distributions?

- The basics
  - Compare their means

\[
d(P_Y, P_Y | X_i) = (\mathbb{E}(Y) - \mathbb{E}(Y | X_i))^2
\]
How can we compare probability distributions?

- The basics
  - Compare their means

\[ d(P_Y, P_{Y|X_i}) = (\mathbb{E}(Y) - \mathbb{E}(Y|X_i))^2 \]  ➔ Sobol!
GENERALIZED GSA

How can we compare probability distributions?

- The basics
  - Compare their means
    \[ d(P_Y, P_{Y|X_i}) = (\mathbb{E}(Y) - \mathbb{E}(Y|X_i))^2 \]  \( \Rightarrow \) Sobol!

- The f-divergence family
  \[ d_f(P_Y \| P_{Y|X_i}) = \int f \left( \frac{p_Y(y)}{p_{Y|X_i}(y)} \right) p_{Y|X_i}(y) dy \]
GENERALIZED GSA

How can we compare probability distributions?

- The basics
  - Compare their means
    \[ d(P_Y, P_{Y|X_i}) = (\mathbb{E}(Y) - \mathbb{E}(Y|X_i))^2 \]  \( \Rightarrow \) Sobol!

- The f-divergence family
  \[ d_f(P_Y || P_{Y|X_i}) = \int f \left( \frac{p_Y(y)}{p_{Y|X_i}(y)} \right) p_{Y|X_i}(y) dy \]

  \[ S_i^f = \int f \left( \frac{p_Y(y)p_{X_i}(x)}{p_{X_i,Y}(x, y)} \right) p_{X_i,Y}(x, y) dxdy \]  D. 2014
GENERALIZED GSA

How can we compare probability distributions?

- The basics
  - Compare their means
    \[ d(P_Y, P_{Y|X_i}) = (\mathbb{E}(Y) - \mathbb{E}(Y|X_i))^2 \]  
    \( \Rightarrow \) Sobol!

- The f-divergence family
  - Includes as particular cases
    \[ d_f(P_Y || P_{Y|X_i}) = \int f \left( \frac{p_Y(y)}{p_{Y|X_i}(y)} \right) p_{Y|X_i}(y) dy \]

\[ S_i^{\text{TV}} = \int |p_Y(y) - p_{Y|X_i=x}(y)| p_X(x) dx dy \]
\[ S_i^{\text{KL}} = \int p_{Y|X_i=x}(y) \ln \left( \frac{p_{Y|X_i=x}(y)}{p_Y(y)} \right) p_X(x) dx dy \]

Borgonovo 2007  
Kraskov et al. 2001
GENERALIZED GSA

How can we compare probability distributions?

- The basics
  - Compare their means
  
  \[ d(P_Y, P_{Y|X_i}) = (\mathbb{E}(Y) - \mathbb{E}(Y|X_i))^2 \]  
  \[ \Rightarrow \text{Sobol!} \]

- The f-divergence family

\[ d_f(P_Y || P_{Y|X_i}) = \int f \left( \frac{p_Y(y)}{p_{Y|X_i}(y)} \right) p_{Y|X_i}(y) dy \]

- Maximum Mean Discrepancy (MMD) or Integral Probability Metrics (IPMs)
GENERALIZED GSA

» Maximum Mean Discrepancy

$$\text{MMD}(P, Q; F) := \sup_{f \in F} \left[ \mathbb{E}_P f(x) - \mathbb{E}_Q f(x) \right]$$

» The distance is zero iif the probability distributions are equal

- $F = \text{bounded continuous functions (Dudley metric)}$
- $F = \text{functions with bounded variations (Kolmogorov metric)}$
- $F = \text{Lipschitz bounded functions (Earth mover’s distance – Wasserstein metric)}$
GENERALIZED GSA

Distributional indices: advantages

- Account for the whole effect of a parameter on the output distribution and not only on the mean
- Density-based, which means
  - Many methods and codes for estimation
  - As we have seen, several distances can be investigated without any additional cost
GENERALIZED GSA

Distributional indices: advantages

- Account for the whole effect of a parameter on the output distribution and not only on the mean
- Density-based, which means
  - Many methods and codes for estimation
  - As we have seen, several distances can be investigated without any additional cost

Limitations

- Density estimation suffers from the curse of dimensionality
  - If we want to consider outputs which are not scalars, this will be a bottleneck
  - Impossible to compute a total index equivalent in this setting
    - Even low order interactions
- Estimation bias
GENERALIZED GSA

Distributional indices: advantages

- Account for the whole effect of a parameter on the output distribution and not only on
  the mean
- Density-based, which means
  - Many methods and codes for estimation
  - As we have seen, several distances can be investigated without any additional cost

Limitations

- Density estimation suffers from the curse of dimensionality
  - If we want to consider outputs which are not scalars, this will be a bottleneck
  - Impossible to compute a total index equivalent in this setting
    - Even low order interactions
- Estimation bias
- No decomposition into main effects, interactions, ...
  - Interpretation is problematic
GENERALIZED GSA

Distributional indices: advantages

- Account for the whole effect of a parameter on the output distribution and not only on the mean
- Density-based, which means
  - Many methods and codes for estimation
  - As we have seen, several distances can be investigated without any additional cost

Limitations

- Density estimation suffers from the curse of dimensionality
  - If we want to consider outputs which are not scalars, this will be a bottleneck
  - Impossible to compute a total index equivalent in this setting
    - Even low order interactions
- Estimation bias
- No decomposition into main effects, interactions, ...
  - Interpretation is problematic
- A possible point of view: RKHS embedding of probability distributions
OUTLINE

- Context

- Generalized GSA
  - Distances between probability distributions
  - RKHS embedding
  - Orthogonal decompositions

- Conclusion & Perspectives
RKHS EMBEDDING

\[ \mathcal{M}^+_1 \]

\( P \)

\( Q \)
\[ d(P, Q) = \sup_{A \in \Sigma} |P(A) - Q(A)| \]

\[ \text{TV} = \frac{1}{2} \int_{\Omega} |f_P - f_Q|d\mu \]
\[ d(P, Q) = \int_{\Omega} f_P \log(f_P / f_Q) d\mu \]

KL

\[ M_{1+} \]

\( P \)

\( Q \)
Other point of view: represent a probability distribution with some features
Other point of view: represent a probability distribution with some features
Other point of view: represent a probability distribution with some features
The dissimilarity between probability distributions is measured through the distance between their representation in the feature space.
Question: which feature?
RKHS EMBEDDING

\[ \mathcal{M}_1^+ \]

Feature Space

\[ \mathcal{F} \]

Dissimilarity measured only through the means

\[ \mu_P = \mathbb{E}_P[X] \]

\[ \mu_Q = \mathbb{E}_Q[X] \]
RKHS EMBEDDING

Two Gaussians with different means

Gretton 2012

Ok ✓
 RKHS EMBEDDING

Two Gaussians with different variances

NOK

Gretton 2012
RKHS EMBEDDING

Two Gaussians with different variances

\[ P_X \quad Q_X \]

\[ \begin{align*}
    X & \quad P_X \\
    -6 & \quad 0.4 \\
    -4 & \quad 0.35 \\
    -2 & \quad 0.3 \\
     0 & \quad 0.25 \\
     2 & \quad 0.2 \\
     4 & \quad 0.15 \\
     6 & \quad 0.1 \\
\end{align*} \]

Densities of feature \( X^2 \)

\[ P_X \quad Q_X \]

\[ \begin{align*}
    X^2 & \quad P_X \\
    10^{-1} & \quad 1.4 \\
    10^0 & \quad 1.2 \\
    10^1 & \quad 1.0 \\
    10^2 & \quad 0.8 \\
\end{align*} \]

\( \checkmark \) OK

\( \times \) NOK

Gretton 2012
RKHS EMBEDDING

\[ \mathcal{M}_1^+ \]

- \( P \)
- \( Q \)

\[ \mathcal{F} \]

Feature Space

\[ \mu_P = [\mathbb{E}_P[X], \mathbb{E}_P[X^2]] \]

\[ \mu_Q = [\mathbb{E}_Q[X], \mathbb{E}_Q[X^2]] \]

Dissimilarity measured only through means & variances
RKHS EMBEDDING

Gaussian and Laplace densities

\[ P_X \]
\[ Q_X \]

NOK

Gretton 2012
RKHS EMBEDDING

General setting: take a feature map

$$\phi : \Omega \rightarrow \mathcal{F}$$
RKHS EMBEDDING

General setting: take a feature map

$$\phi : \Omega \rightarrow \mathcal{F}$$
Instead of choosing the feature map, make it implicit and assume that the feature space is a RKHS with a given kernel

\[ k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}} \]
Instead of choosing the feature map, make it implicit and assume that the feature space is a RKHS with a given kernel

\[ k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}} \]
RKHS EMBEDDING

In practice

- Choose the kernel
- How can the distance be computed in the feature space?
RKHS EMBEDDING

**In practice**
- Choose the kernel
- How can the distance be computed in the feature space?

\[
\|\mu_P - \mu_Q\|_\mathcal{F}^2 = \langle \mu_P, \mu_P \rangle_\mathcal{F} + \langle \mu_Q, \mu_Q \rangle_\mathcal{F} - 2\langle \mu_P, \mu_Q \rangle_\mathcal{F} \\
= \ldots \\
= \mathbb{E}_{X \sim P, X' \sim P}[k(X, X')] + \mathbb{E}_{Y \sim Q, Y' \sim Q}[k(Y, Y')] - 2\mathbb{E}_{X \sim P, Y \sim Q}[k(X, Y)]
\]

Standard reproducing RKHS property

- Distance which involves only the kernel
  - Kernel trick in action

- Several nice papers on the subject
RKHS EMBEDDING: REMEMBER MMD?

Maximum Mean Discrepancy

\[ \text{MMD}(P, Q; F) := \sup_{f \in F} [E_P f(x) - E_Q f(x)] \]

The distance is zero iif the probability distributions are equal

- \( F = \) bounded continuous functions (Dudley metric)
- \( F = \) functions with bounded variations (Kolmogorov metric)
- \( F = \) Lipschitz bounded functions (Earth mover’s distance – Wasserstein metric)
RKHS EMBEDDING: REMEMBER MMD?

- Maximum Mean Discrepancy

\[
\text{MMD}(P, Q; F) \equiv \sup_{f \in F} \left[ \mathbb{E}_P f(x) - \mathbb{E}_Q f(x) \right]
\]

- The distance is zero iif the probability distributions are equal
  - \( F = \) bounded continuous functions (Dudley metric)
  - \( F = \) functions with bounded variations (Kolmogorov metric)
  - \( F = \) Lipschitz bounded functions (Earth mover’s distance – Wasserstein metric)
  - \( F = \) unit ball in a characteristic RKHS (Sriperumbudur et al. 2008)
Maximum Mean Discrepancy in a RKHS

\[
\text{MMD}^2(P, Q; F) = \left( \sup_{f \in F} \left[ \mathbb{E}_P f(x) - \mathbb{E}_Q f(x) \right] \right)^2
\]

\[
= \left( \sup_{f \in F} \langle f, \mu_P - \mu_Q \rangle_F \right)^2
\]

\[
= \| \mu_P - \mu_Q \|^2_F
\]

MMD point of view and feature space point of view are equivalent
RKHS EMBEDDING: TOWARDS GSA

- General framework

\[ S_i = \mathbb{E}_{X_i} \left( d(P_Y, P_Y|X_i) \right) \]

- If we use the MMD distance

\[ d(P_Y, P_Y|X_i) = \text{MMD}^2(P_Y, P_Y|X_i) \]

\[ S_i = \mathbb{E}_{X_i} \left( \text{MMD}^2(P_Y, P_Y|X_i) \right) \]

\[ S_i = \int \Omega k(y, y') \left[ p_Y(y) - p_Y|X_i=x_i(y) \right] \left[ p_Y(y') - p_Y|X_i=x_i(y') \right] p_X(x_i) dy dy' dx_i \]
\[ S_i = \mathbb{E}_{X_i} (\text{MMD}^2(P_Y, P_{Y|X_i})) \]
A few remarks

- You can choose any kernel
- If we want to distinguish probability distributions, we must use a characteristic kernel
  - e.g. Gaussian, exponential
- But in practice you can choose any kernel, including

\[ k(y, y') = \langle y, y' \rangle_{1D} = yy' \]

Feature map is identity
Comparison through means only

Fukumizu et al. (2008)
Sriperumbudur et al. (2008)
RKHS EMBEDDING: TOWARDS GSA

- **A few remarks**
  - You can choose any kernel
  - **If we want to distinguish probability distributions, we must use a characteristic kernel**
    - e.g. Gaussian, exponential
  - But in practice you can choose any kernel, including

\[
k(y, y') = \langle y, y' \rangle^{1D} = yy'
\]

Feature map is identity
Comparison through means only

\[
\mathbb{E} \left( \text{MMD}^2(P_Y, P_Y | X_i) \right) = \text{Var}(\mathbb{E}(Y | X_i))
\]

Unnormalized Sobol index

Fukumizu et al. (2008)
Sriperumbudur et al. (2008)
A few remarks

- You can choose any kernel
- If we want to distinguish probability distributions, we must use a characteristic kernel
  - e.g. Gaussian, exponential

- But in practice you can choose any kernel, including

  \[ k(y, y') = \langle y, y' \rangle^{1D} = yy' \]

- This is thus a natural extension of Sobol

Fukumizu et al. (2008)
Sriperumbudur et al. (2008)
A few remarks

- You can choose any kernel
- If we want to distinguish probability distributions, we must use a characteristic kernel
  - e.g. Gaussian, exponential

- But in practice you can choose any kernel, including

  \[ k(y, y') = \langle y, y' \rangle_{1D} = yy' \]

  Feature map is identity
  Comparison through means only

- This is thus a natural extension of Sobol
- Question: where does the normalizing constant come from?
  - Sobol-Hoeffding decomposition!
  - Can we have the same?
 RKHS EMBEDDING: DECOMPOSITION I

⇒ Re-write the Sobol-Hoeffding decomposition

\[
\text{Var}(Y) = \sum_{u \subseteq \{1, \ldots, d\}, u \neq \emptyset} g_u
\]

\[
g_u = \sum_{v \subseteq u} (-1)^{|u| - |v|} \text{Var}\left(\mathbb{E}(Y|X_v)\right)
\]
Theorem (D. 2016)

\[
\mathbb{E} \left( \text{MMD}^2 \left( P_Y | X_{1:d}, P_Y \right) \right) = \sum_{u \subseteq \{1, \ldots, p\}, u \neq \emptyset} g_u
\]

\[
g_u = \sum_{v \subseteq u} (-1)^{|u| - |v|} \mathbb{E} \left( \text{MMD}^2 \left( P_Y | X_v, P_Y \right) \right)
\]
Theorem (D. 2016)

\[ \mathbb{E} \left( \text{MMD}^2 \left( P_Y | X_{1:d}, P_Y \right) \right) = \sum_{u \subseteq \{1, \ldots, p\}, u \neq \emptyset} g_u \]

\[ g_u = \sum_{v \subseteq u} (-1)^{|u| - |v|} \mathbb{E} \left( \text{MMD}^2 \left( P_Y | X_v, P_Y \right) \right) \]

MMD sensitivity indices

\[ S_u^{\text{MMD}} = \frac{\sum_{v \subseteq u} (-1)^{|u| - |v|} \mathbb{E} \left( \text{MMD}^2 \left( P_Y | X_v, P_Y \right) \right)}{\mathbb{E} \left( \text{MMD}^2 \left( P_Y | X_{1:d}, P_Y \right) \right)} \]
Alternate interpretation of MMD indices

- If we use a Mercer kernel,

\[ k(y, y') = \sum_{j=1}^{\infty} \Phi_j(y) \Phi_j(y') \]

- As a result, 1\textsuperscript{st} order MMD indices are given by

\[ S_i^{\text{MMD}} = \frac{\sum_{j=1}^{\infty} \text{Var}(\mathbb{E}(\Phi_j(Y)|X_i))}{\sum_{j=1}^{\infty} \text{Var}(\Phi_j(Y))} = \sum_{j=1}^{\infty} \alpha_j S_i^{\text{Sobol}} [\phi_j(Y)] \]

*Linear combination of the Sobol index for a (potential) infinity of transformations of the output (i.e. features)*
RKHS EMBEDDING: ESTIMATION

- Standard MMD estimation

\[
\hat{\text{MMD}}^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} [k(x_i, x_j) - k(x_i, x'_j) - k(x'_i, x_j) + k(x'_i, x'_j)]
\]

\[
\{x_i\}_{i=1}^{n} \sim P, \quad \{x'_i\}_{i=1}^{n} \sim Q
\]
RKHS EMBEDDING: ESTIMATION

- Standard MMD estimation

\[
\hat{\text{MMD}}^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} [k(x_i, x_j) - k(x_i, x'_j) - k(x'_i, x_j) + k(x'_i, x'_j)]
\]

\[
\{x_i\}_{i=1}^{n} \sim P, \ \{x'_i\}_{i=1}^{n} \sim Q
\]

- What about the MMD sensitivity index?

\[
S_i = \mathbb{E}_{X_i} \left( \text{MMD}^2(P_Y, P_Y|X_i) \right)
\]

- Brute-force Monte-Carlo very expensive
- Possible to use Pick & Freeze estimation
- Ongoing investigation of replicated designs to get rid of the input dimension
We can go even further!
 RKHS EMBEDDING: FEATURE SELECTION

⇒ We can go even further!

⇒ Remember the density-based index

\[
S_{i}^{KL} = \int p_{Y|X_i=x(y)} \ln \left( \frac{p_{Y|X_i=x(y)}}{p_{Y}(y)} \right) p_{X_i}(x) \, dx \, dy
\]
RKHS EMBEDDING: FEATURE SELECTION

- We can go even further!

- Remember the density-based index

\[
S_{KL}^i = \int p_{Y|X_i = x}(y) \ln \left( \frac{p_{Y|X_i = x}(y)}{p_Y(y)} \right) p_X(x) dx dy
\]

\[
= \int p_{Y,X_i}(y, x) \ln \left( \frac{p_{Y,X_i}(y, x)}{p_Y(y)p_X(x)} \right) dx dy = I(X_i; Y)
\]

\text{Mutual Information}

- In this case, the sensitivity index is a dependence measure between random variables
From a broad perspective, a dependence measure compares the joint distribution and the product of the marginals

- If close, the variables are dependent
- How do we compare the joint distribution and the product of the marginals?
From a broad perspective, a dependence measure compares the joint distribution and the product of the marginals

- If close, the variables are dependent
- How do we compare the joint distribution and the product of the marginals?

\[
\text{MMD}^2(P_{Y,X}, P_Y P_X) = \left( \sup_{f \in F} \left[ \mathbb{E}_{P_{XY}} f(x, y) - \mathbb{E}_{P_X P_Y} f(x, y) \right] \right)^2
\]

\[
= \left\| \mu_{P_{XY}} - \mu_{P_X P_Y} \right\|_{\mathcal{F} \times \mathcal{G}}^2
\]

\[= \text{HSIC}(X, Y) \quad \text{Hilbert-Schmidt Independence Criterion}
\]

Gretton 2005
RKHS EMBEDDING: FEATURE SELECTION

→ HSIC estimation from a sample of the joint distribution

\[ \hat{\text{HSIC}}(X, Y) = \frac{1}{n^2} \text{trace}(KHLH) \]

\[
[K]_{ij} = k_X(x_i, x_j) \quad [L]_{ij} = k_Y(y_i, y_j) \quad [H]_{ij} = \delta_{ij} - \frac{1}{n}
\]

\[ \{(x_i, y_i)\}_{i=1}^n \sim P_{XY} \]

→ Several feature selection techniques based on this measure
  - Song et al. (2007a,b,c), Balasubramanian et al. (2013), Yamada et al. 2013
- In a GSA context, just rank the input parameters according to their HSIC value with the output
  - A normalization inspired by SRC is proposed in D. (2014)

- Good screening properties
  - At a very low computational cost (~ 100, independent of the input dimension)
In a GSA context, just rank the input parameters according to their HSIC value with the output
- A normalization inspired by SRC is proposed in D. (2014)

Good screening properties
- At a very low computational cost (~ 100, independent of the input dimension)

Would be great if we could use this measure as a sensitivity index
- With particular case the MMD indices
- With a decomposition
- Link between feature selection and GSA
RKHS EMBEDDING: DECOMPOSITION II

Theorem (D. 2016)

\[
\text{HSIC} (Y, X_{1:d}) = \sum_{u \subseteq \{1, \ldots, p\}, u \neq \emptyset} g_u
\]

\[
g_u = \sum_{v \subseteq u} (-1)^{|u| - |v|} \text{HSIC} (Y, X_v)
\]

If the kernel on each input satisfies

\[
\int_X k_X(x, x') \, dP_X(x) = 1
\]

\[
k_X = 1 + k_X^0
\]

\[
k_X^0(x, x') = k(x, x') - \frac{\int_X k(x, x') \, dP_X(x') \int_X k(x, x') \, dP_X(x)}{\int_{X \times X} k(x, x') \, dP_X(x) \, dP_X(x')}
\]

\[
k_X^0(x, x') = k(x, x') - \int_X k(x, x') \, dP_X(x') - \int_X k(x, x') \, dP_X(x)
\]
Theorem (D. 2016)

$$\text{HSIC} (Y, X_{1:d}) = \sum_{u \subseteq \{1, \ldots, p\}, u \neq \emptyset} g_u$$

$$g_u = \sum_{v \subseteq u} (-1)^{|u| - |v|} \text{HSIC} (Y, X_v)$$

$$S_{u}^{\text{HSIC}} = \frac{\sum_{v \subseteq u} (-1)^{|u| - |v|} \text{HSIC} (Y, X_v)}{\text{HSIC} (Y, X_{1:d})}$$
More remarks

- You have to choose a kernel for each input and output
- **If we want to detect independence, we must use a characteristic kernel**
  - e.g. Gaussian, exponential
- The decomposition holds for a centered-like kernel
  - Actually same assumption for the ANOVA-kernel of Durrande et al. (2013)

Fukumizu et al. (2008)
Sriperumbudur et al. (2008)
More remarks

- You have to choose a kernel for each input and output
- If we want to detect independence, we must use a characteristic kernel
  - e.g. Gaussian, exponential
- The decomposition holds for a centered-like kernel
  - Actually same assumption for the ANOVA-kernel of Durrande et al. (2013)
- This is a natural extension again

\[ k_X(x, x') \rightarrow \delta(x, x') \]

**Example:**

\[ k_X(x, x') = \frac{1}{\sqrt{2\pi a^2}} \exp \left( -\frac{1}{2a^2} (x - x')^2 \right), \quad a \rightarrow 0 \]

\[ S_u^{\text{HSIC}} \rightarrow S_u^{\text{MMD}} \]
More remarks

- You have to choose a kernel for each input and output
- **If we want to detect independence, we must use a characteristic kernel**
  - e.g. Gaussian, exponential
- The decomposition holds for a centered-like kernel
  - Actually same assumption for the ANOVA-kernel of Durrande et al. (2013)
- This is a natural extension again

\[
k_X(x, x') \to \delta(x, x')
\]

\[
\begin{align*}
S_{u}^{\text{HSIC}} & \longrightarrow S_{u}^{\text{MMD}} \\
S_{u}^{\text{MMD}} & = S_{u}^{\text{Sobol}}
\end{align*}
\]

ex: \( k_X(x, x') = \frac{1}{\sqrt{2\pi a^2}} \exp \left( \frac{1}{2a^2} (x - x')^2 \right), \ a \to 0 \)
The RKHS point of view comes with a huge literature and dedicated kernels

- If your inputs or outputs are vectors, curves, texts, images, timeseries, DNA sequences, probability distributions, ... there is a kernel available
  - We then have a generic GSA framework which can handle them, with a decomposition into main effects and interactions
Example 1: migration of strontium 90 in a storage site (Marthe testcase CEA)

- **Inputs**
  - 20 geological parameters

- **Outputs**
  - Strontium concentration at 10 observation wells
  - 2D maps of concentration (64x64=4096 pixels)

Marrel et al. 2011
RKHS EMBEDDING: LET’S PLAY WITH KERNELS

Gaussian kernel in 10 dimensions (wells)

PCA kernel on 2D maps

D. 2014
Example 2: Optimization

\[
\min_{x \in \mathcal{X}} f(x) \\
\text{s.t. } h(x) \leq 0
\]

- Fact: the number of local minima grows exponentially w.r.t. dimensionality
  - And exploring the feasibility domain also suffers from the curse of dimensionality

- Question: how can we adapt GSA tools for optimization problems?
  - Usually, perform GSA on all objectives & constraints
    - Relevant variables are the intersection
    - (Pray for a dimension reduction …)
  - Or use multi-output GSA
Example 2: Optimization

\[
\min_{x \in \mathcal{X}} f(x)
\]

s.t. \( h(x) \leq 0 \)

- Problem with standard approach: **basic GSA is not adapted to the goal**
  - It will detect which inputs impact the mean level of the objective and the constraints

- What we really want:
  - **Identify which inputs lead us to a feasible region & a low value of the objective**
Example 2: Optimization

Dixon-Price function

One variable has a sensitivity index equal to 1

But in order to find the global minimum, the other variable is essential!
RKHS EMBEDDING: LET’S PLAY WITH KERNELS

Example 2: Optimization

\[ f((X_1, X_3)|X_2) \]

Ishigami function

Variable 2 has a significant Sobol index

But whatever its value, the regions of minimum and maximum can only be found by using the other 2 variables!

Master’s internship of Adrien Spagnol @ IRT SystemX
Example 2: Optimization & dedicated kernels

Master's internship of Adrien Spagnol @ IRT SystemX
The RKHS point of view comes with a huge literature and dedicated kernels

- If your inputs or outputs are vectors, curves, texts, images, timeseries, DNA sequences, probability distributions, … there is a kernel available
  - We then have a generic GSA framework which can handle them, with a decomposition into main effects and interactions

- And you can recover previously studied sensitivity indices with particular kernels
RKHS EMBEDDING: LET’S PLAY WITH KERNELS

The RKHS point of view comes with a huge literature and dedicated kernels

- If your inputs or outputs are vectors, curves, texts, images, timeseries, DNA sequences, probability distributions, … there is a kernel available
  - We then have a generic GSA framework which can handle them, with a decomposition into main effects and interactions

- And you can recover previously studied sensitivity indices with particular kernels

\[ k_X(x, x') \rightarrow \delta(x, x') \]

ex: \[ k_X(x, x') = \frac{1}{\sqrt{2\pi a^2}} \exp\left(\frac{1}{2a^2}(x - x')^2\right), \ a \rightarrow 0 \]

\[ S_u^{\text{HSIC}} \rightarrow S_u^{\text{MMD}} \]

\[ k_Y(y, y') = yy' \]

\[ S_u^{\text{MMD}} = S_u^{\text{Sobol}} \]
The RKHS point of view comes with a huge literature and dedicated kernels

- If your inputs or outputs are vectors, curves, texts, images, timeseries, DNA sequences, probability distributions, ... there is a kernel available
  - We then have a generic GSA framework which can handle them, with a decomposition into main effects and interactions

- And you can recover previously studied sensitivity indices with particular kernels

\[
\begin{align*}
    k_{\mathcal{X} \_ i}(x, x') & \rightarrow \delta(x, x') \\
    k_{\mathcal{X} \_ i}(x, x') & \rightarrow \delta'(x, x') \\
    k_{\mathcal{Y}}(y, y') & = yy'
\end{align*}
\]
The RKHS point of view comes with a huge literature and dedicated kernels

- If your inputs or outputs are vectors, curves, texts, images, timeseries, DNA sequences, probability distributions, … there is a kernel available
  - We then have a generic GSA framework which can handle them, with a decomposition into main effects and interactions

- And you can recover previously studied sensitivity indices with particular kernels

\[
\begin{align*}
    k_{\chi_{-i}}(x, x') &\rightarrow \delta(x, x') \\
    k_{\chi_i}(x, x') &\rightarrow \delta'(x, x') \\
    k_{\gamma}(y, y') &= yy'
\end{align*}
\]

\[
S_{i}^{\text{HSIC}} \rightarrow \int_{\Omega} \left( \frac{\partial \eta(x)}{\partial x_i} \right)^2 \, dx
\]

We recover the 1st order DGSM indices!
CONCLUSION

→ New sensitivity index which generalizes GSA through the use of kernels

\[
S_u^{\text{HSIC}} = \sum_{v \subseteq u} (-1)^{|u| - |v|} \frac{\text{HSIC}(Y, X_v)}{\text{HSIC}(Y, X_{1:d})}
\]

- In theory, this is a density-based index: better measure of the influence than a mere mean
- Limiting cases include Sobol & DGSM
- Decomposition into main effects & interactions: interpretation is possible
- Built upon a feature selection technique: the frontier between screening methods and quantitative approaches may disappear
I honestly think there is potential there, but
- Extensive benchmark studies are still needed
  - In particular kernels for curves, 3D objects, …
- Application to optimization

From a theoretical perspective
- Investigate the links with ANOVA-kernels
- See if we can recover other sensitivity indices as particular cases
- Use replicated designs for MMD indices estimation

Should be soon available in the R package sensitivity
CONCLUSION

COME SEE US AT THE NEXT GDR ANNUAL MEETING

SAFRAN – MASSY
22/03 – 24/03 2017
REFERENCES


REFERENCES


