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**

CONTEXT

➔ 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

Computer code

$$Y = \eta(X_1, \ldots, X_d)$$

Output

Input parameters

CONTEXT

➔ Two points of view

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

$$S_i = \frac{\sigma^2_{X_i}}{\mathrm{Var}(Y)} \left(\frac{\partial \eta(X)}{\partial X_i} |_{X=X_0} \right)^2$$

  - *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)

CONTEXT

➔ Global Sensivity Analysis (GSA) – 2 families of methods

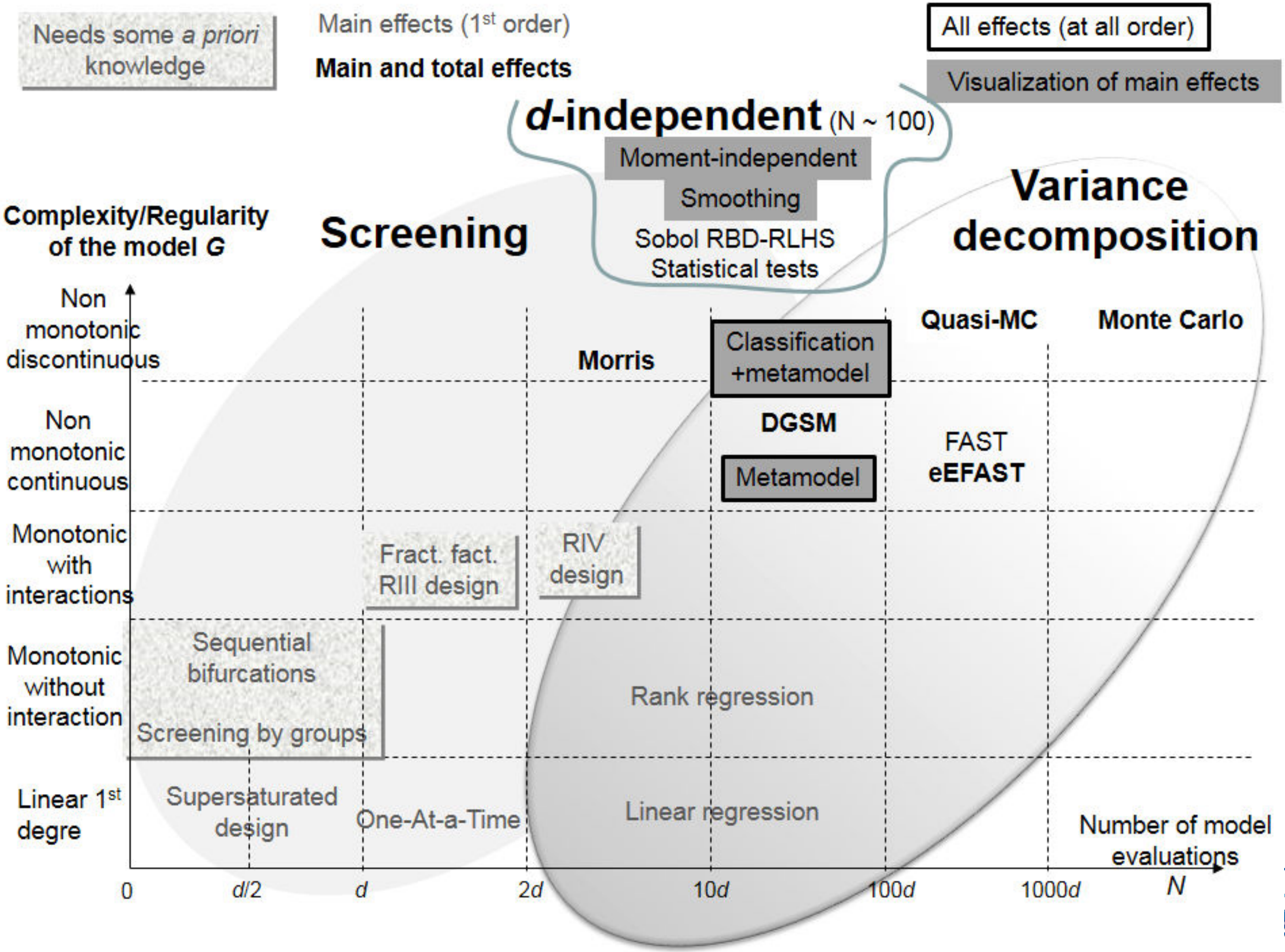
- Screening methods
  - Standard DOEs
  - Sequential bifurcation, …
  - Morris

$$n \approx d/2 - 10d$$

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

$$n \approx 2d - 10^4 d$$

CONTEXT



Iooss & Saltelli 2015

SAFRAN
AEROSPACE · DEFENCE · SECURITY

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:

$$\begin{aligned}
\eta_0 &= \mathbb{E}(Y) \\
\eta_i(X_i) &= \mathbb{E}(Y|X_i) - \mathbb{E}(Y) \\
\eta_{i,j}(X_i, X_j) &= \mathbb{E}(Y|X_i, X_j) - \mathbb{E}(Y|X_i) - \mathbb{E}(Y|X_j) + \mathbb{E}(Y) \\
&\ldots
\end{aligned}$$

CONTEXT

➔ GSA – Focus on Sobol indices

- By orthogonality

$$\mathrm{Var}(\eta(X)) = \sum_{i=1}^{d} \mathrm{Var}(\eta_i(X_i)) + \sum_{1 \leq i < j \leq d} \mathrm{Var}(\eta_{i,j}(X_i, X_j)) + \ldots + \mathrm{Var}(\eta_{1,\ldots,d}(X_1, \ldots, X_d))$$

  - The total variance is decomposed into pieces involving main effects, 2nd order interactions, and so on
  - => Possibility to define the sensitivity index of a groupt of input parameters

$$S_I(X_I) = \frac{\mathrm{Var}(\eta_I(X_I))}{\mathrm{Var}(\eta(X))}$$

$$S_i(X_i) = \frac{\mathrm{Var}(\mathbb{E}(Y|X_i))}{\mathrm{Var}(Y)}$$ Main effect

$$S_i^T(X_i) = \sum_{I \supseteq i} S_I$$ Total effect

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 practive, a first screening step is necessary

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

CONTEXT

➔ Limitations

Take Home Message I
Generalized GSA

- 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 practive, a first screening step is necessary

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

CONTEXT

➔ Limitations

Take Home Message I
Generalized GSA

- 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 practive, a first screening step is necessary

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

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

CONTEXT

➔ Limitations

Take Home Message I
Generalized GSA

- 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 practive, a first screening step is necessary

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

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)

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

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

$$\theta_i(x) = \arg\min_{\theta} \mathbb{E}(\psi(Y;\theta)|X_i = x)$$

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)

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

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

$$\theta_i(x) = \arg\min_{\theta} \mathbb{E}(\psi(Y;\theta)|X_i = x)$$

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

$$S_i^{TV} = \int |p_Y(y) - p_{Y|X_i=x}(y)| p_{X_i}(x) dx dy$$

Borgonovo 2007

$$S_i^{KL} = \int p_{Y|X_i=x}(y) \ln\left(\frac{p_{Y|X_i=x}(y)}{p_Y(y)}\right) p_{X_i}(x) dx dy$$

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

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

Baucells and Borgonovo 2013
D. 2014

  - 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

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

Baucells and Borgonovo 2013
D. 2014

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

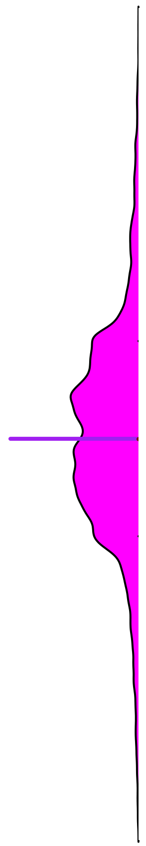
- Toy example

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

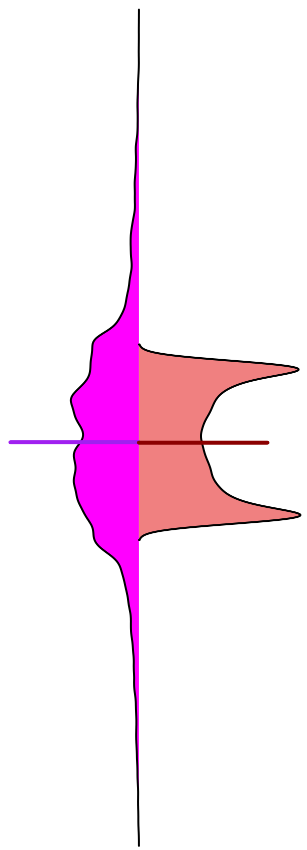
$$X_1, X_2, X_3, X_4 \sim U(-\pi, \pi)$$

Ishigami function with
dummy variable

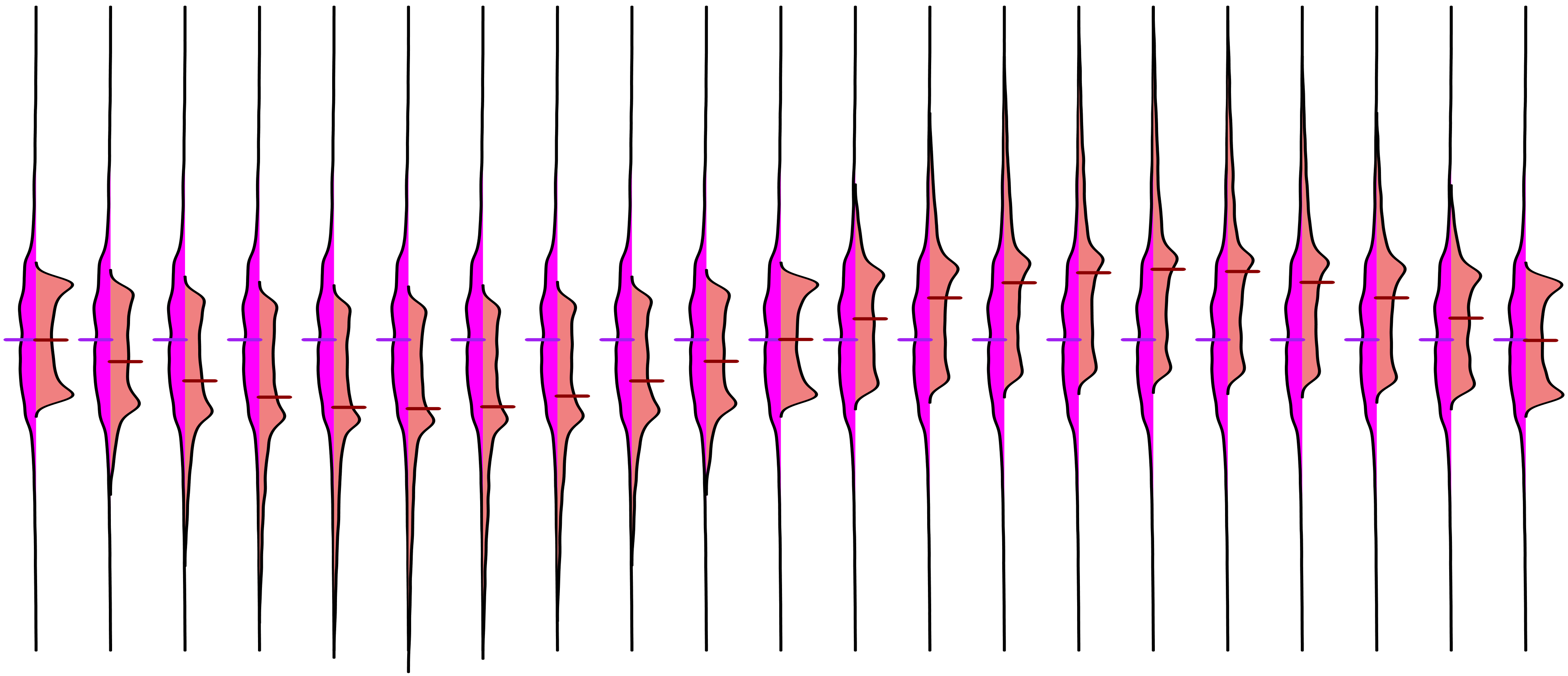
GENERALIZED GSA

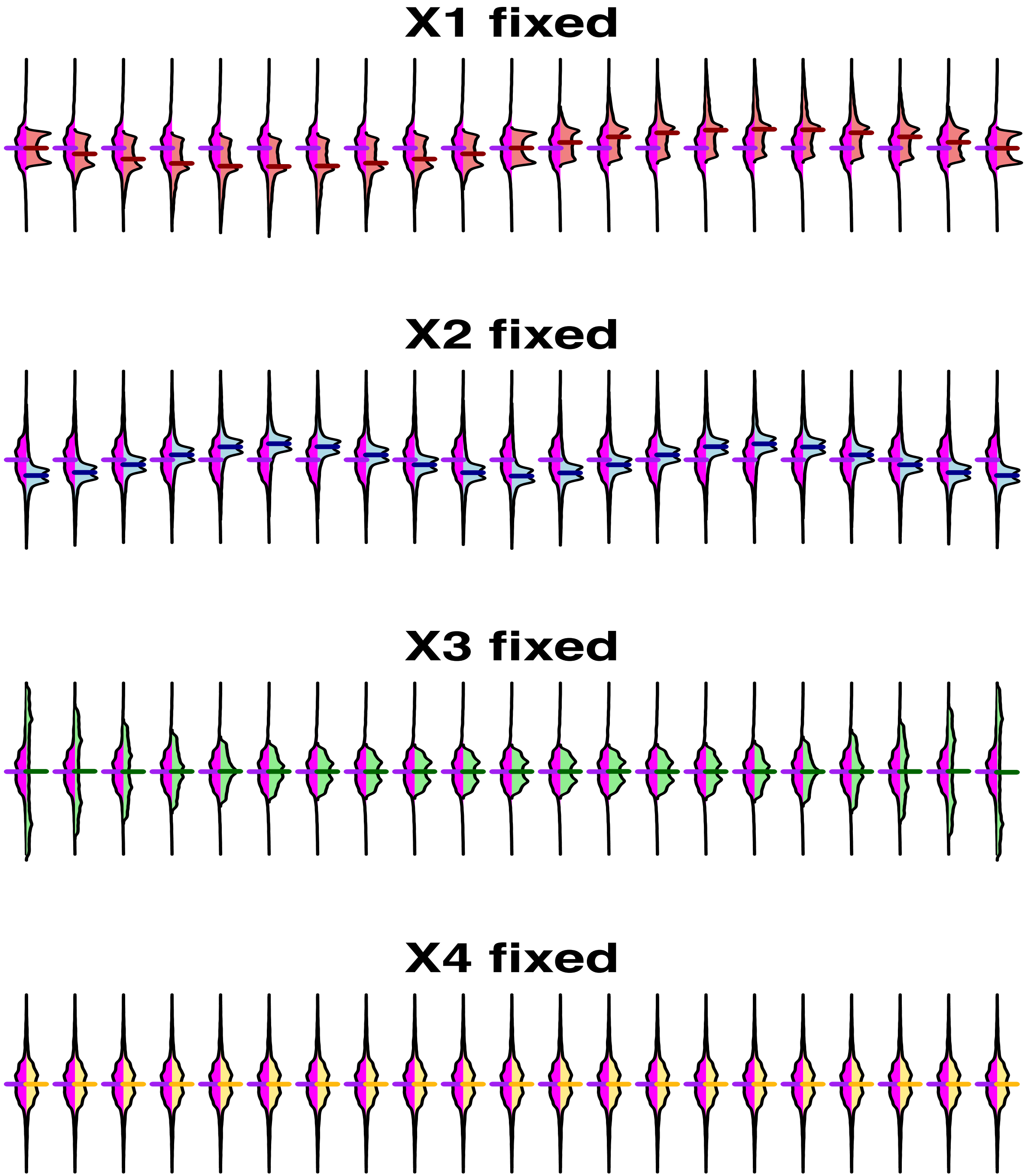


GENERALIZED GSA



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

$$d(P_Y, P_{Y|X_i}) = \left(\mathbb{E}(Y) - \mathbb{E}(Y|X_i)\right)^2$$

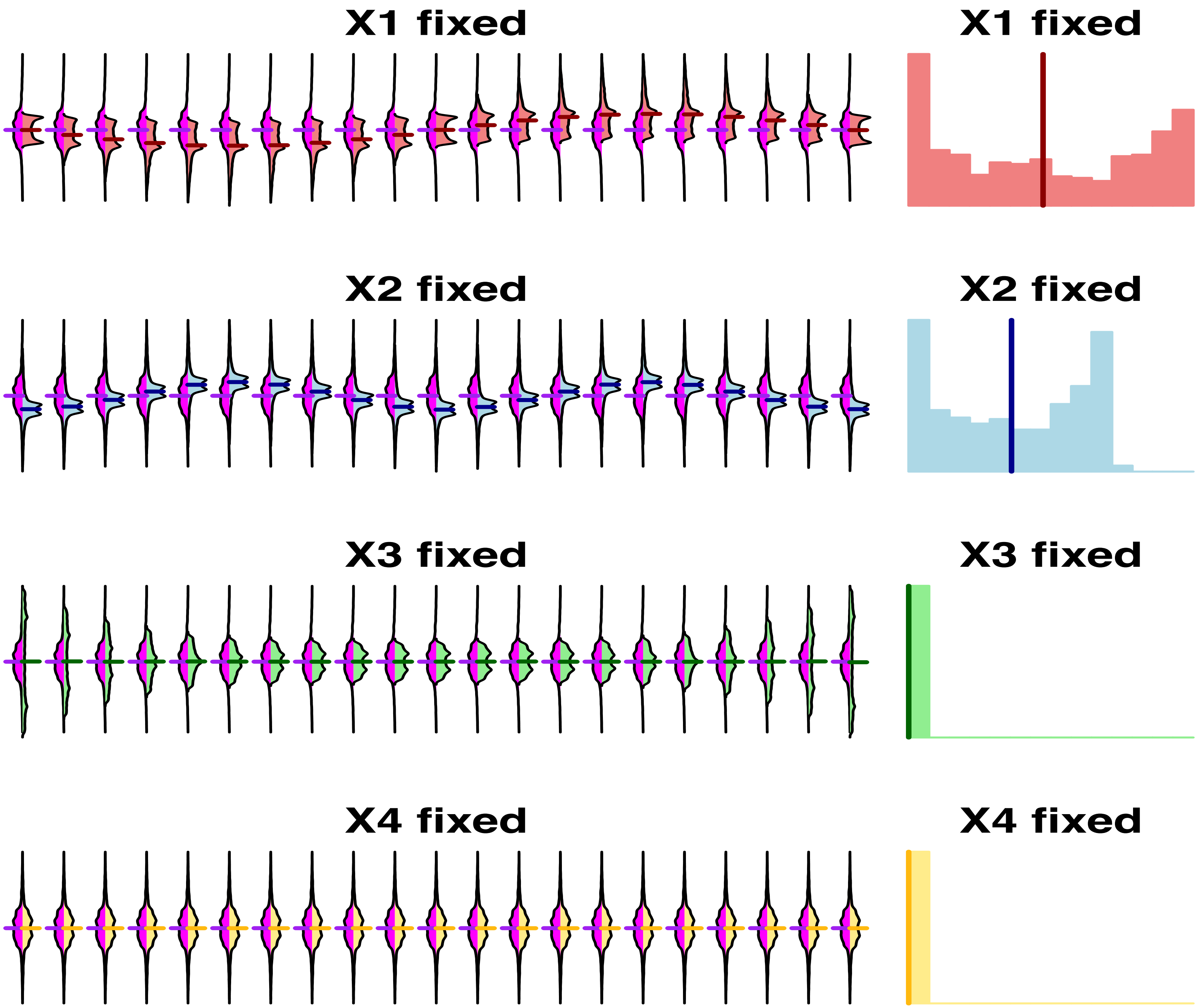
GENERALIZED GSA

➔ How can we compare probability distributions ?

- The basics
  - Compare their means

$$d(P_Y, P_{Y|X_i}) = (\mathbb{E}(Y) - \mathbb{E}(Y|X_i))^2$$

➔ Sobol !


X1 fixed
X1 fixed
X2 fixed
X2 fixed
X3 fixed
X3 fixed
X4 fixed
X4 fixed

GENERALIZED GSA

➔ How can we compare probability distributions ?

- The basics
  - Compare their means

$$d(P_Y, P_{Y|X_i}) = (\mathbb{E}(Y) - \mathbb{E}(Y|X_i))^2$$

➔ 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

$$d(P_Y, P_{Y|X_i}) = (\mathbb{E}(Y) - \mathbb{E}(Y|X_i))^2$$

➔ 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

SAFRAN
AEROSPACE · DEFENCE · SECURITY

GENERALIZED GSA

➔ How can we compare probability distributions ?

- The basics
  - Compare their means

$$d(P_Y, P_{Y|X_i}) = \left(\mathbb{E}(Y) - \mathbb{E}(Y|X_i)\right)^2$$

➔ Sobol !

- The f-divergence family

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

  - Includes as particular cases

-

$$S_i^{TV} = \int |p_Y(y) - p_{Y|X_i=x}(y)| p_{X_i}(x) dx dy$$

Borgonovo 2007

$$S_i^{KL} = \int p_{Y|X_i=x}(y) \ln\left(\frac{p_{Y|X_i=x}(y)}{p_Y(y)}\right) p_{X_i}(x) dx dy$$

Kraskov et al. 2001

GENERALIZED GSA

➔ How can we compare probability distributions ?

- The basics
  - Compare their means

$$d(P_Y, P_{Y|X_i}) = (\mathbb{E}(Y) - \mathbb{E}(Y|X_i))^2$$

➔ 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**

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

➔ **The distance is zero iif the probability distributions are equal**

- F = bounded continuous functions (Dudley metric)
- F = functions with bounded variations (Kolmogorov metric)
- F = 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

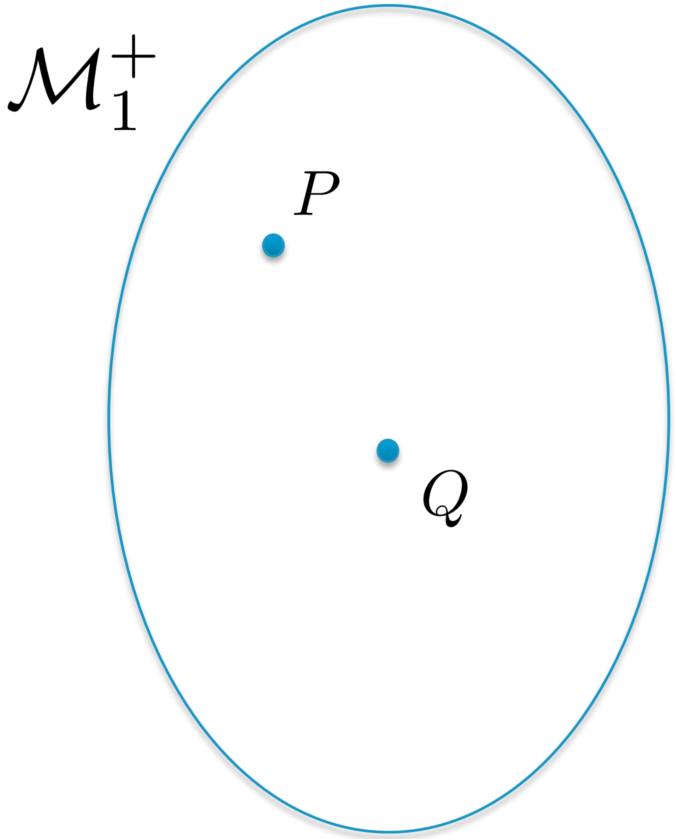
OUTLINE

- Context

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

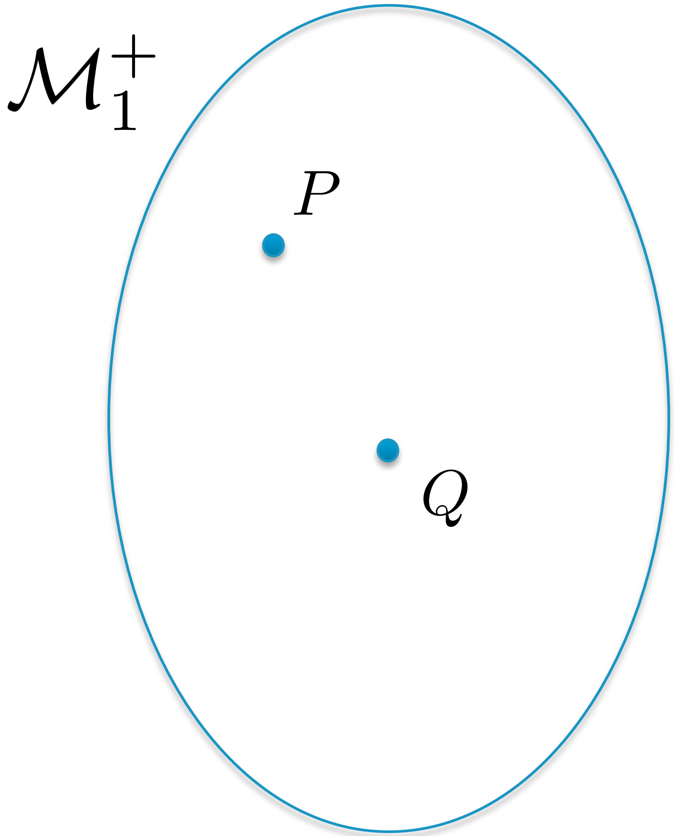
- Conclusion & Perspectives

RKHS EMBEDDING



SAFRAN AEROSPACE · DEFENCE · SECURITY

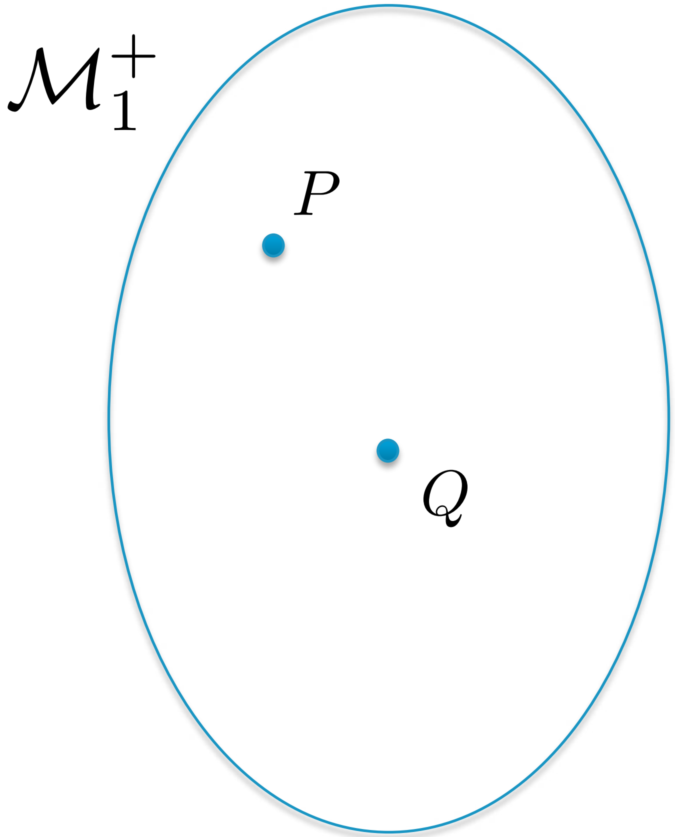
RKHS EMBEDDING



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

TV $$= \frac{1}{2} \int_{\Omega} |f_P - f_Q| d\mu$$

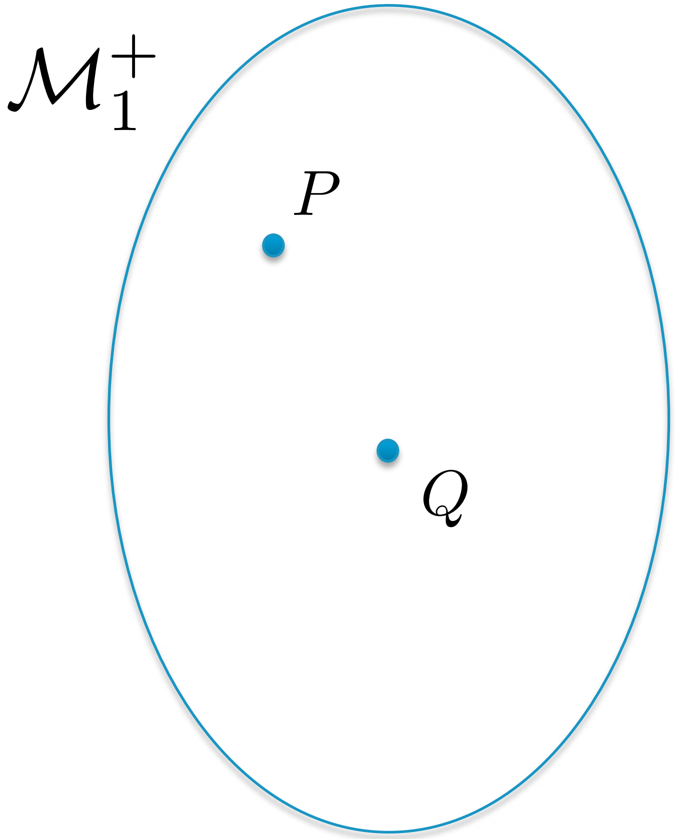
RKHS EMBEDDING



$$d(P,Q) = \int_\Omega f_P log(f_P/f_Q) d\mu$$

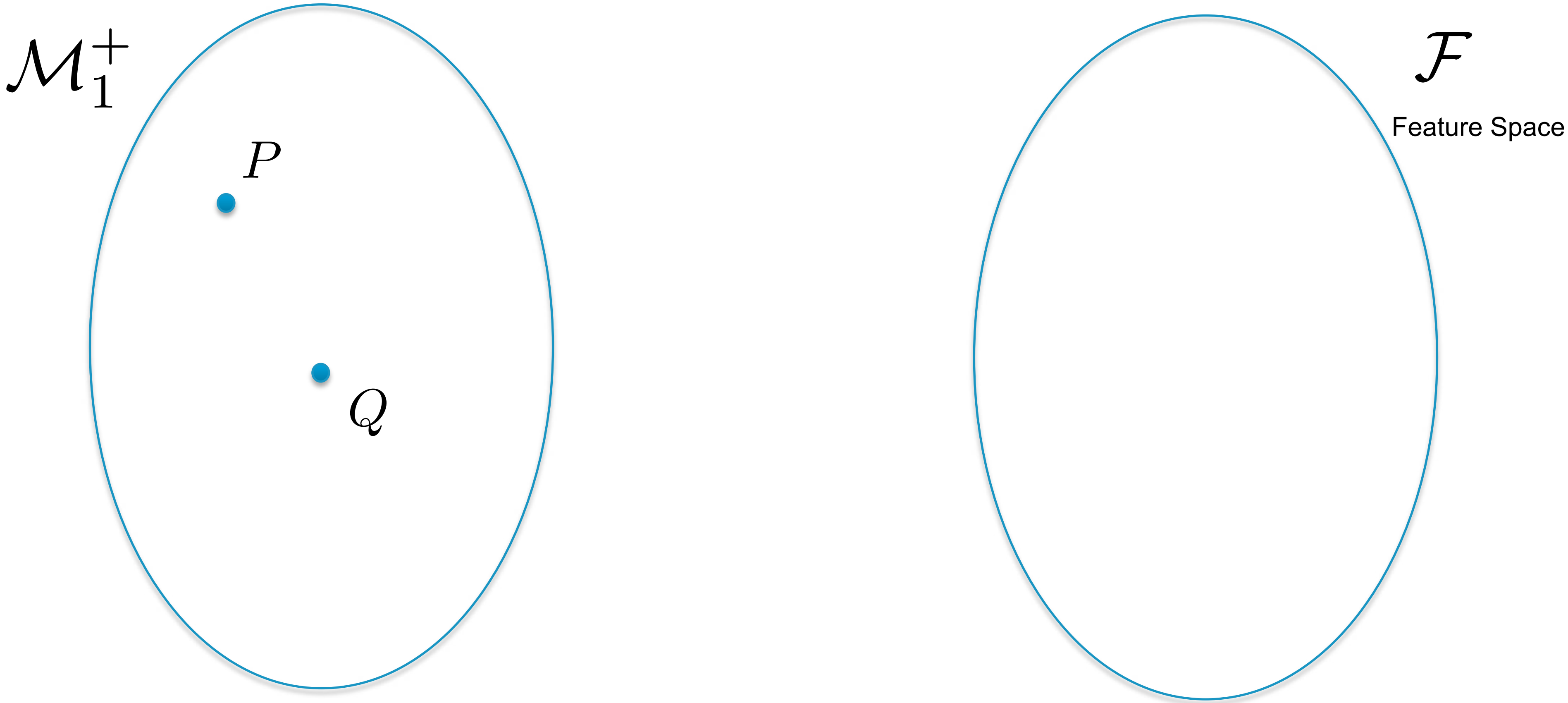
KL

RKHS EMBEDDING



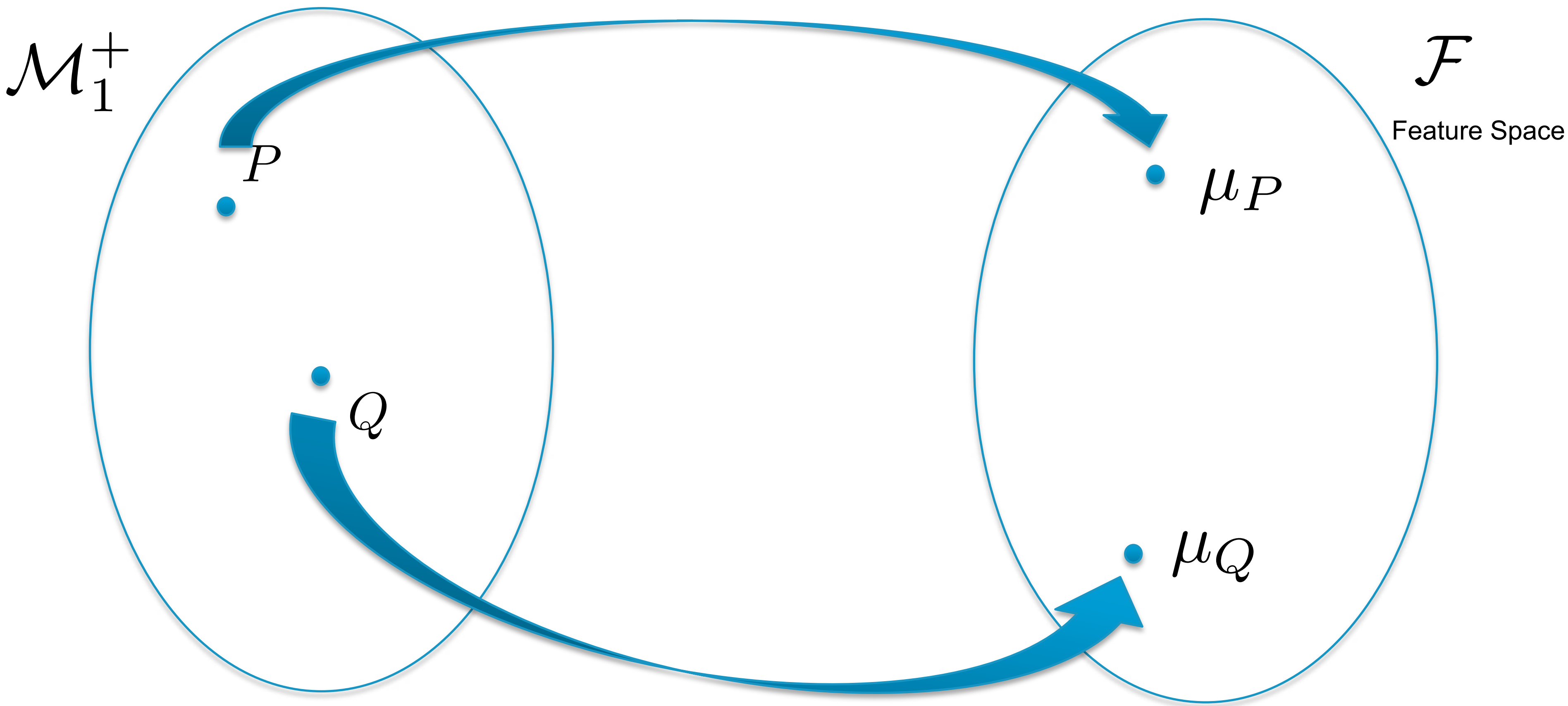
Other point of view: represent a probability distribution with some features

RKHS EMBEDDING



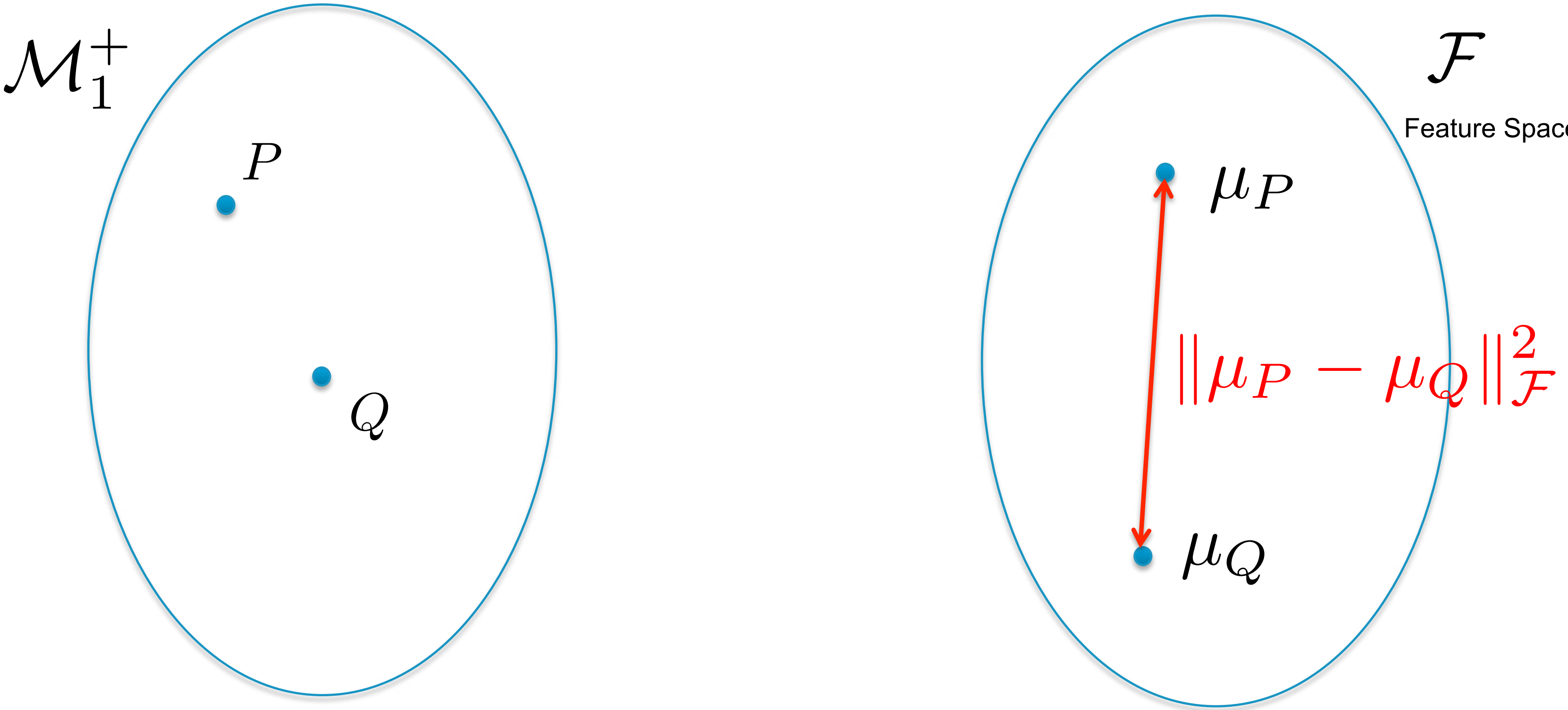
Other point of view: represent a probability distribution with some features

RKHS EMBEDDING



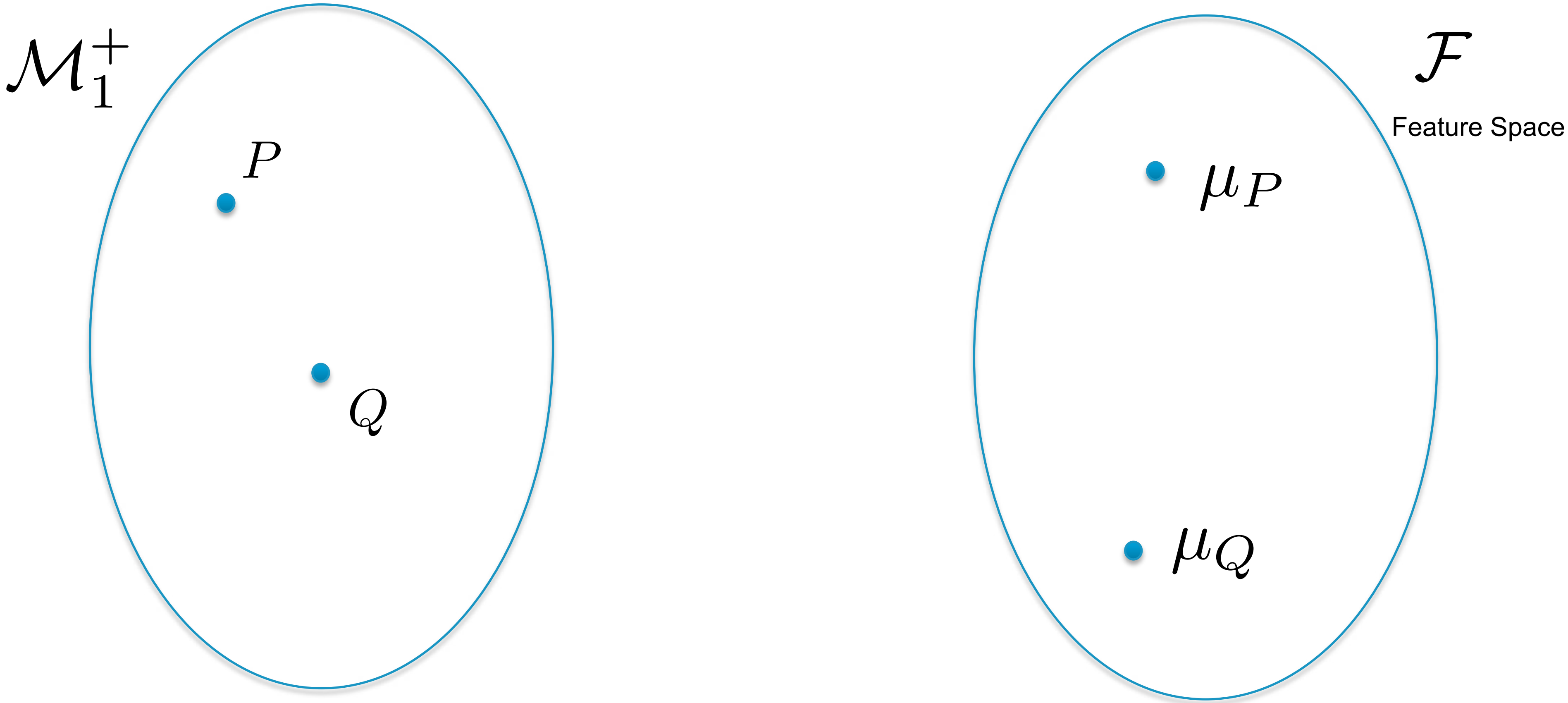
Other point of view: represent a probability distribution with some features

RKHS EMBEDDING



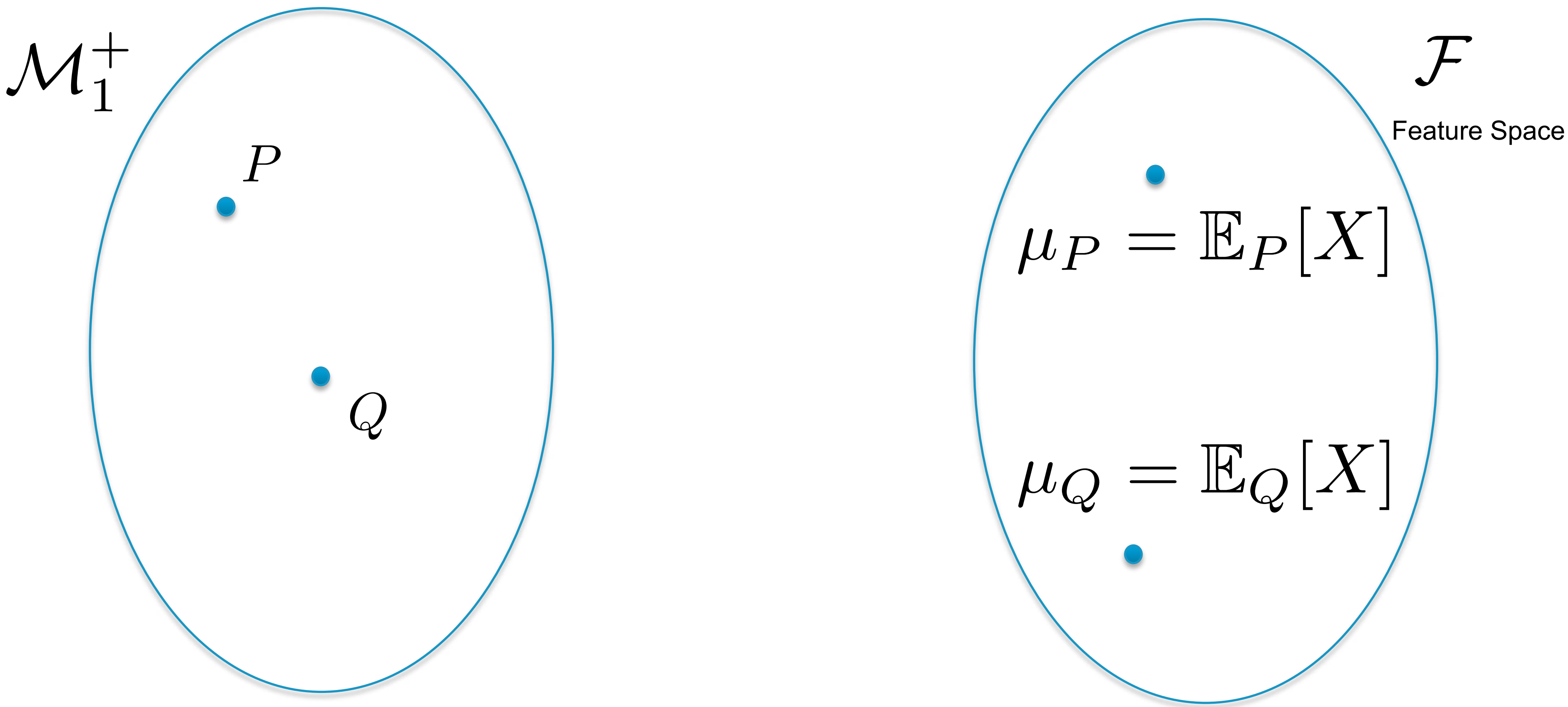
The dissimilarity between probability distributions is measured through the distance between their representation in the feature space

RKHS EMBEDDING



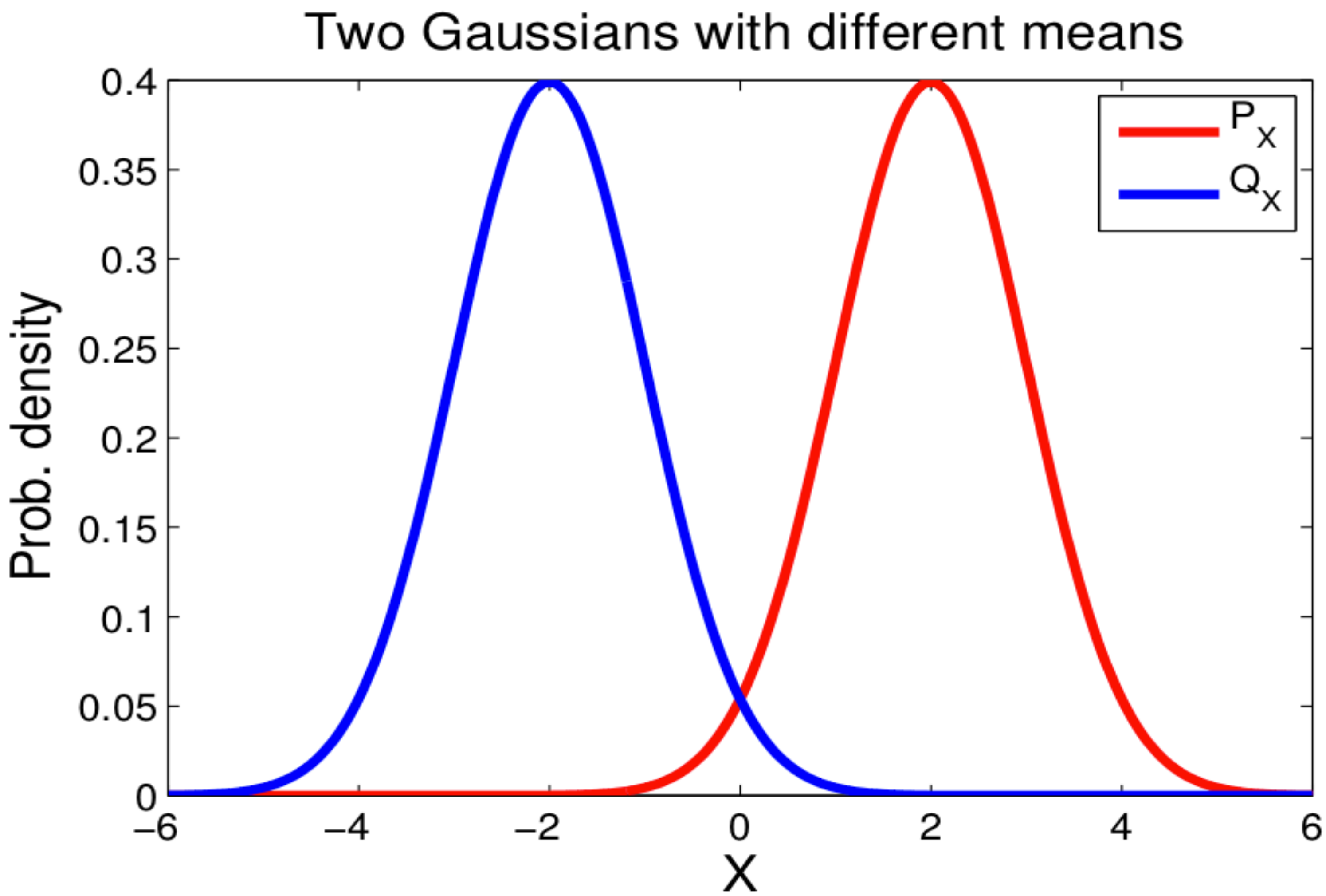
Question: which feature ?

RKHS EMBEDDING



Dissimilarity measured only through the means

RKHS EMBEDDING

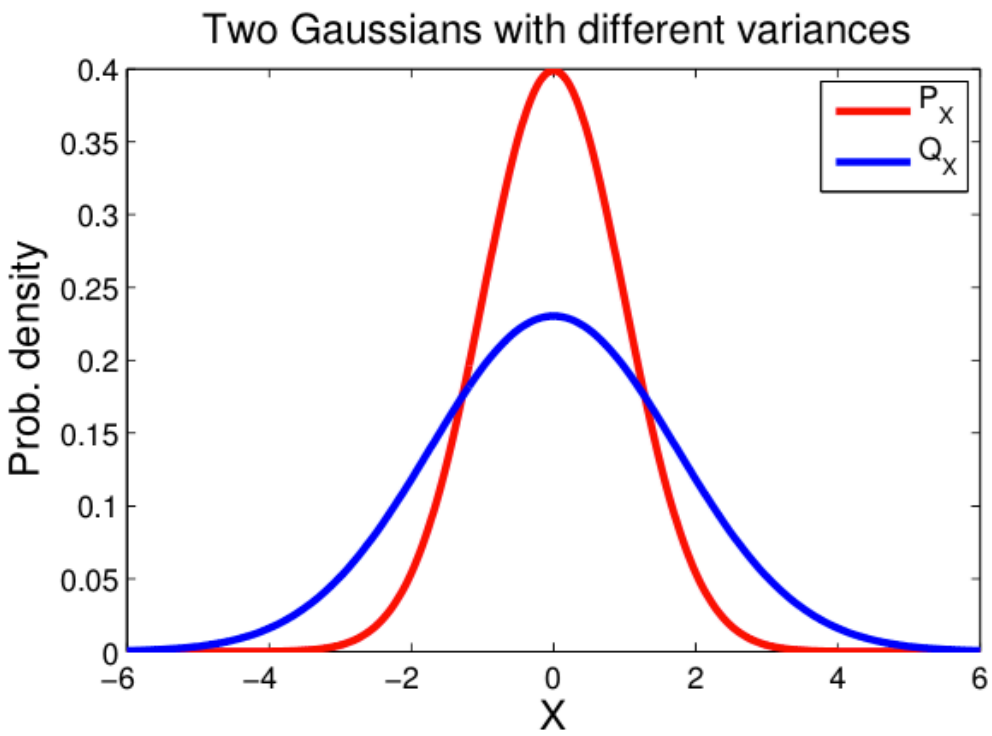


OK

✓

Gretton 2012

RKHS EMBEDDING

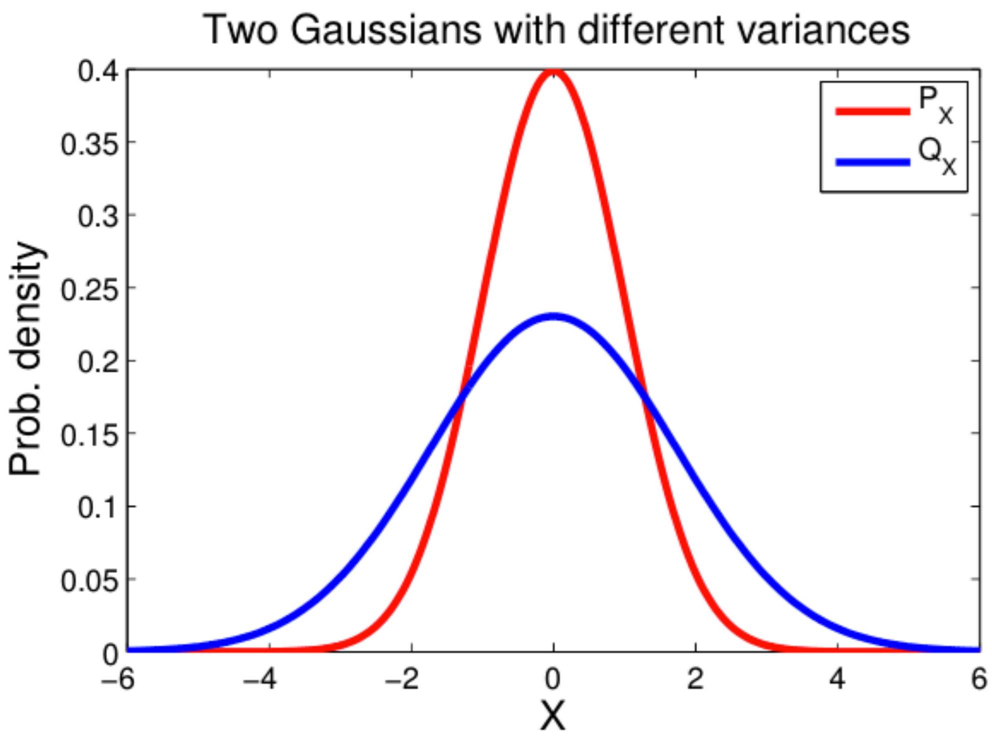


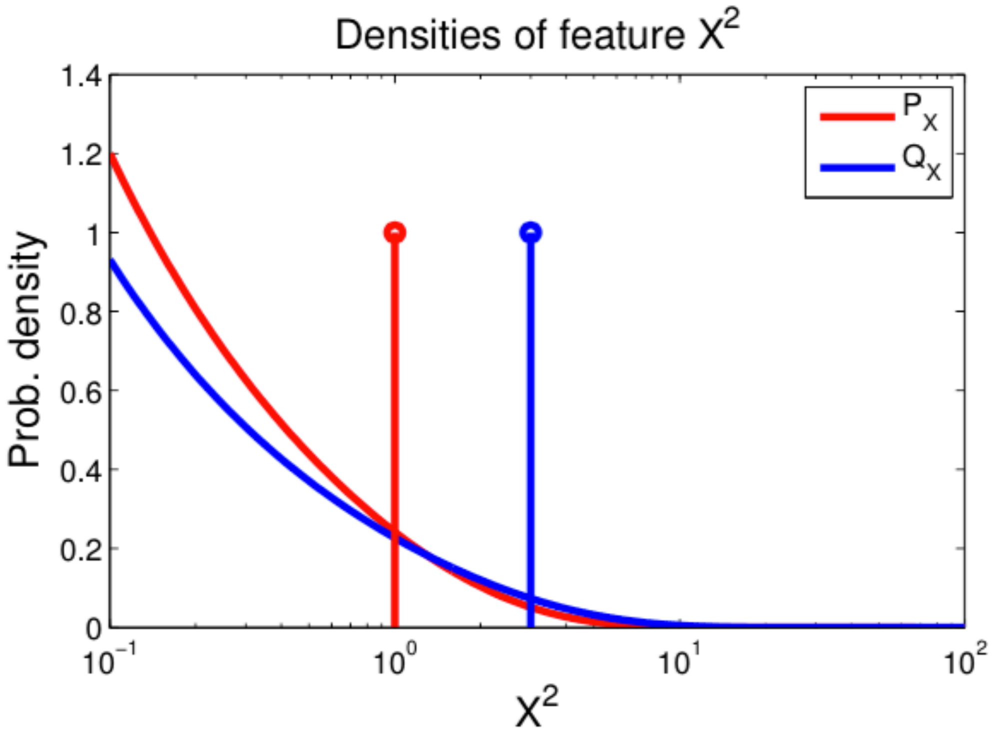
NOK

✗

Gretton 2012

RKHS EMBEDDING



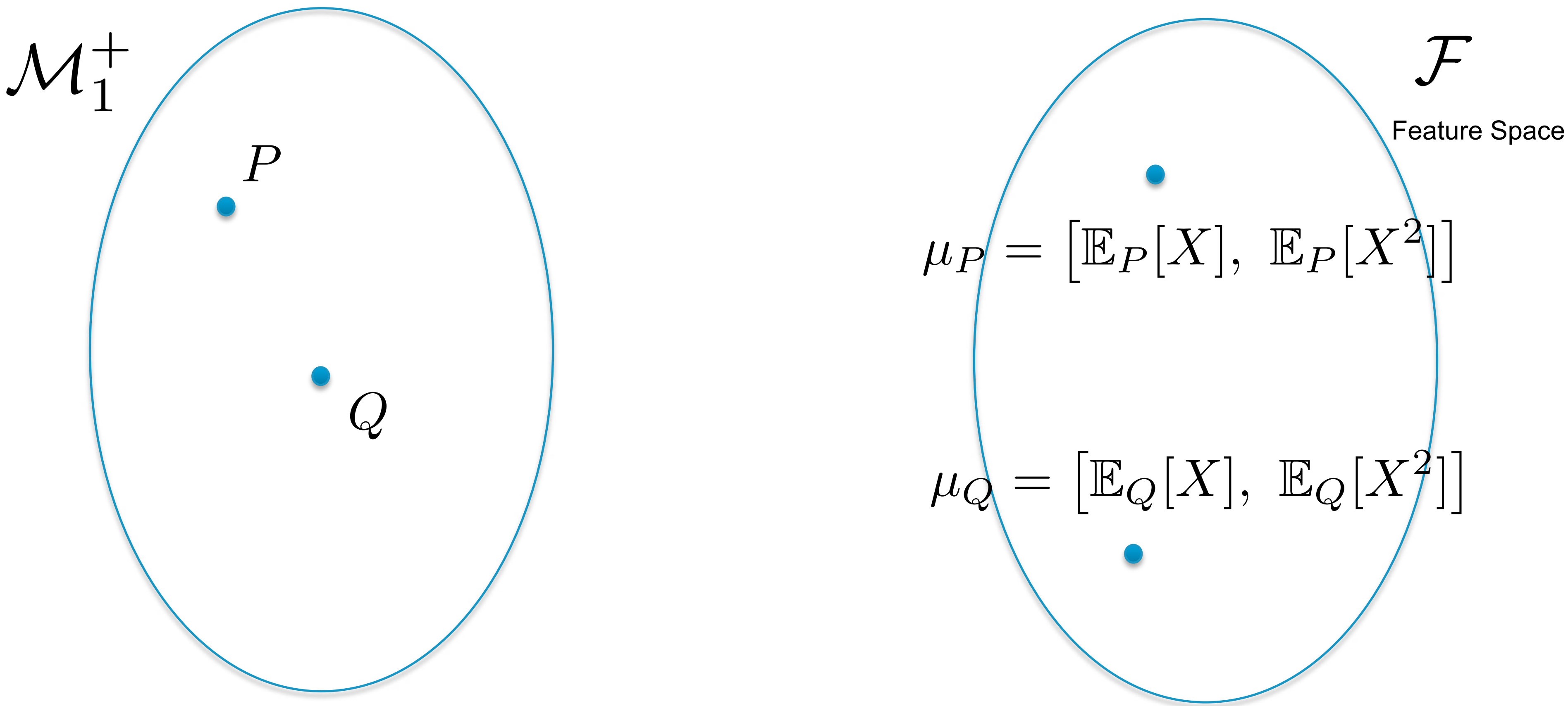


NOK ✗

OK ✓

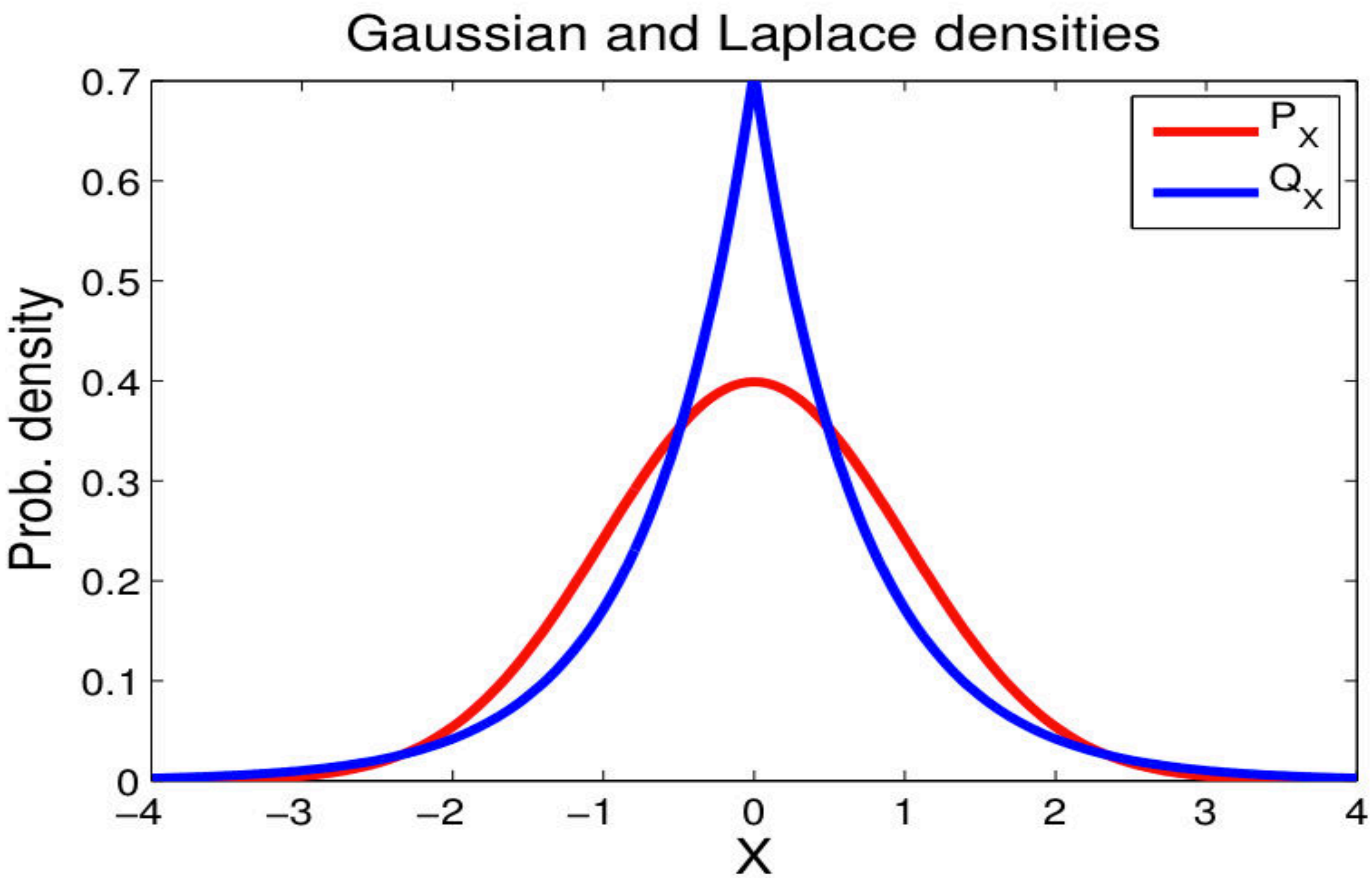
Gretton 2012

RKHS EMBEDDING



Dissimilarity measured only through means & variances

RKHS EMBEDDING

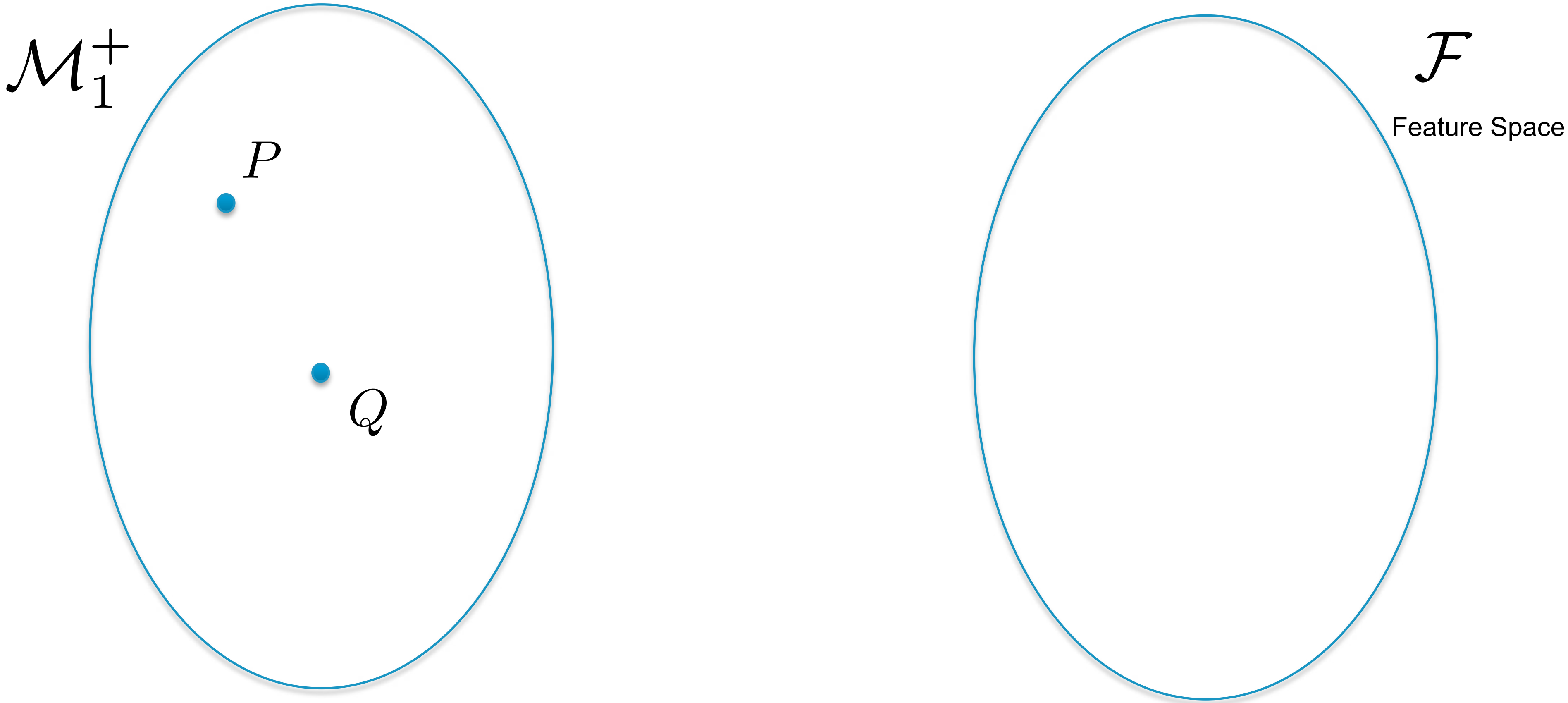


NOK

✗

Gretton 2012

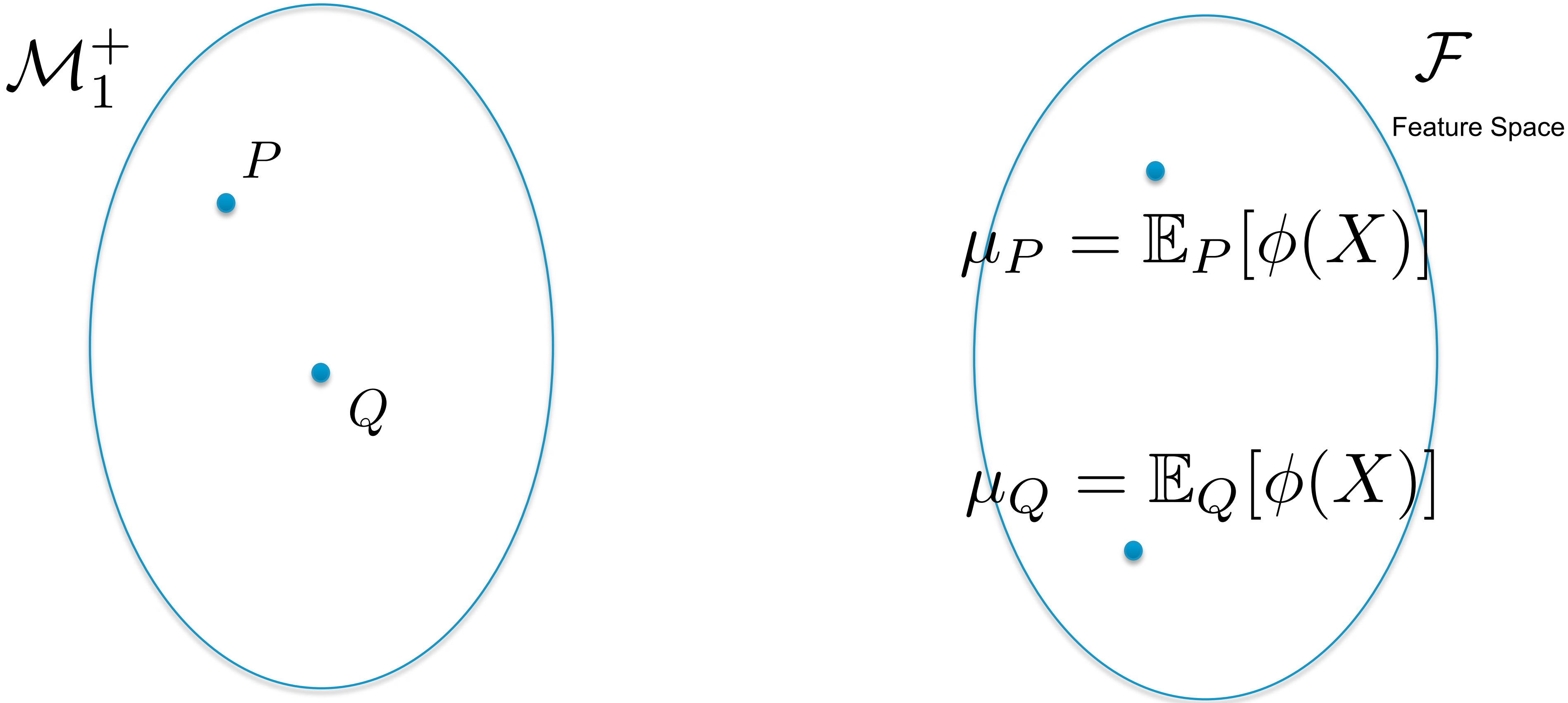
RKHS EMBEDDING



General setting: take a feature map

$$\phi : \ \Omega \to \mathcal{F}$$

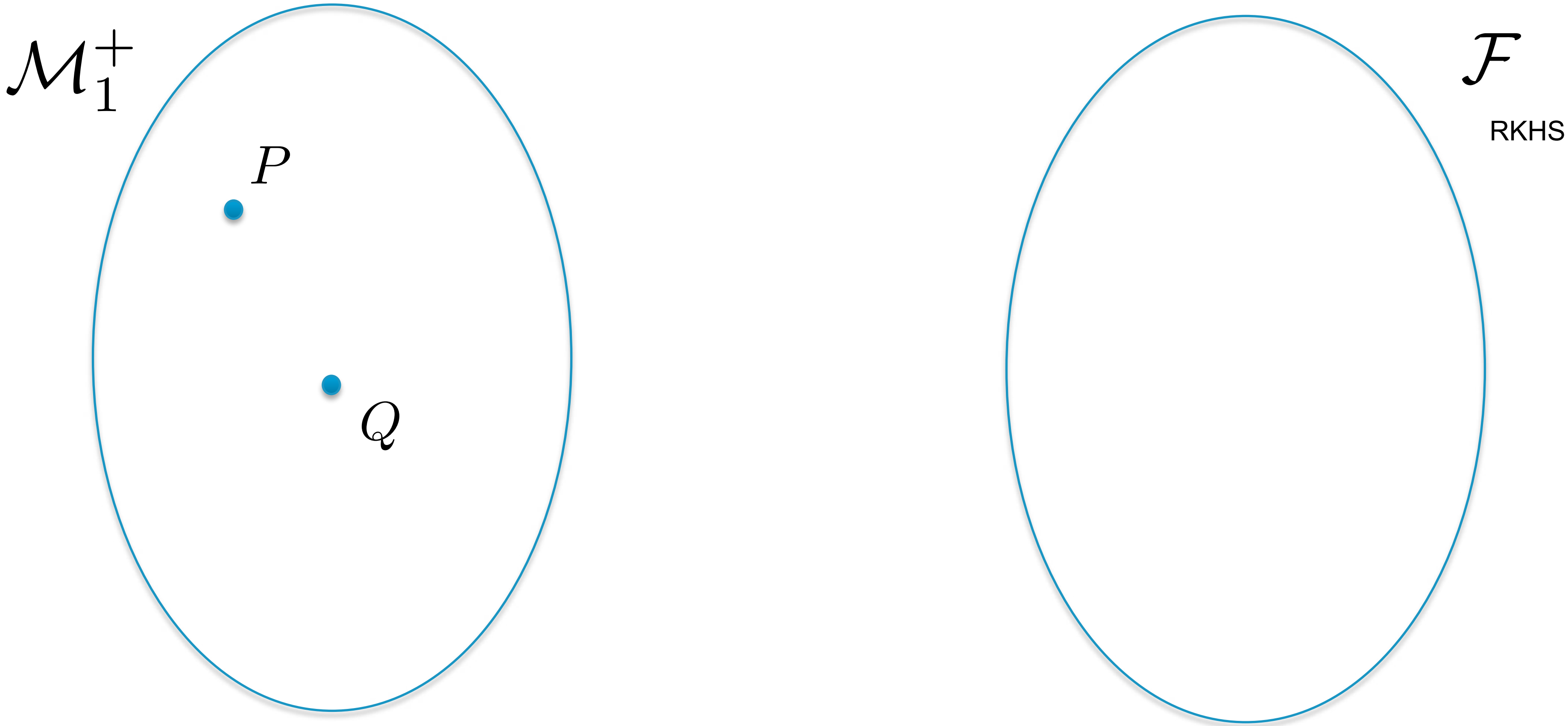
RKHS EMBEDDING



General setting: take a feature map

$$\phi : \ \Omega \to \mathcal{F}$$

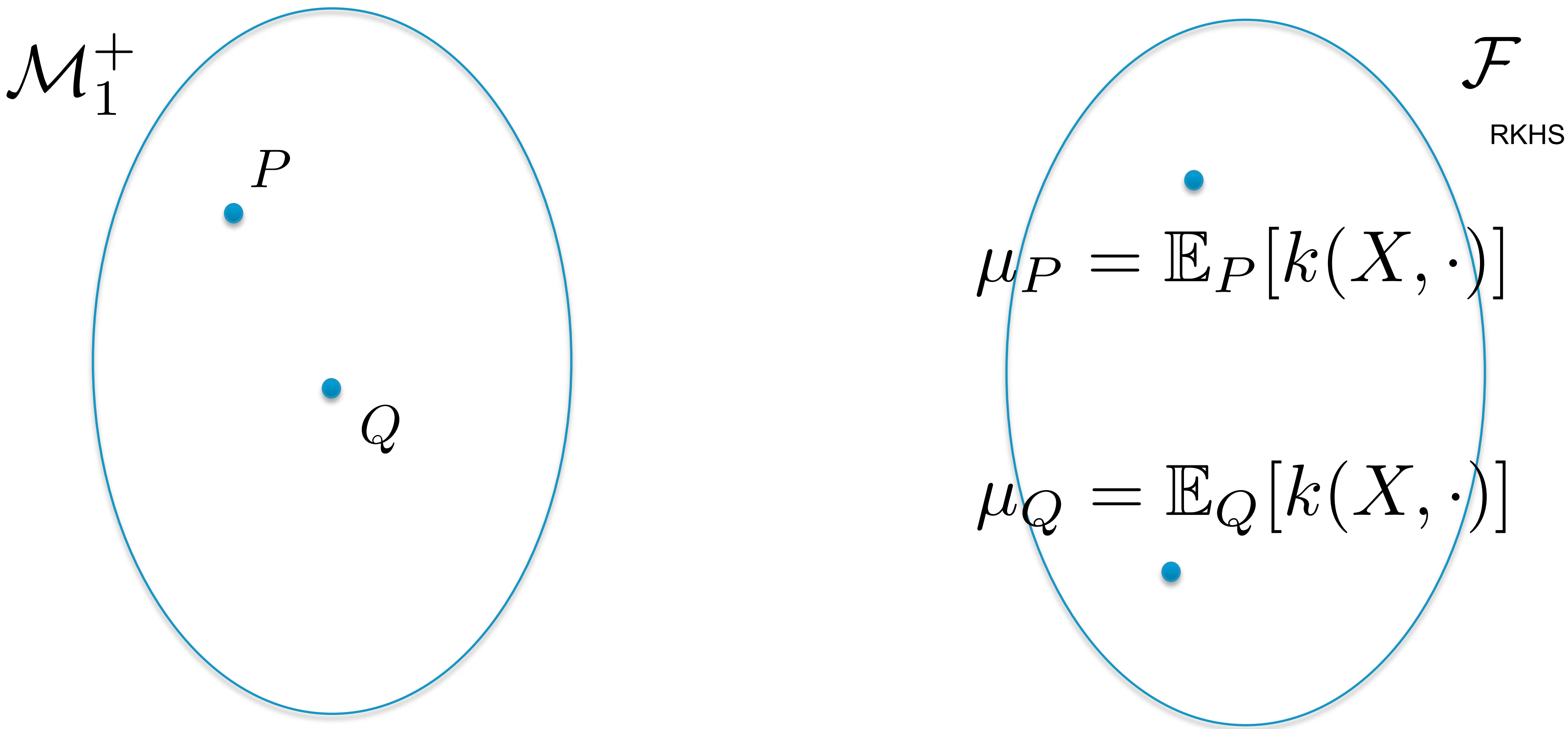
RKHS EMBEDDING



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



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 ?

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

Standard reproducing RKHS property

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

- Several nice papers on the subject
  - Smola et al. 2007, Song 2008, Song et al. 2009

RKHS EMBEDDING: REMEMBER MMD ?

- **Maximum Mean Discrepancy**

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

- **The distance is zero iif the probability distributions are equal**
  - F = 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**

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

➔ **The distance is zero iif the probability distributions are equal**

- F = 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)

RKHS EMBEDDING: REMEMBER MMD ?

➔ Maximum Mean Discrepancy in a RKHS

$$
\begin{aligned}
\mathrm{MMD}^2(P,Q;F) &= \left(\sup_{f\in F}\left[\mathbb{E}_P f(x) - \mathbb{E}_Q f(x)\right]\right)^2 \\
&= \left(\sup_{f\in F}\langle f, \mu_P - \mu_Q\rangle_{\mathcal{F}}\right)^2 \\
&= \|\mu_P - \mu_Q\|^2_{\mathcal{F}}
\end{aligned}
$$

$$\|h\|_{\mathcal{F}} = \sup_{f\in F}\langle f, h\rangle_{\mathcal{F}}$$

MMD point of view and feature space point of view are equivalent

RKHS EMBEDDING: TOWARDS GSA

➔ General framework

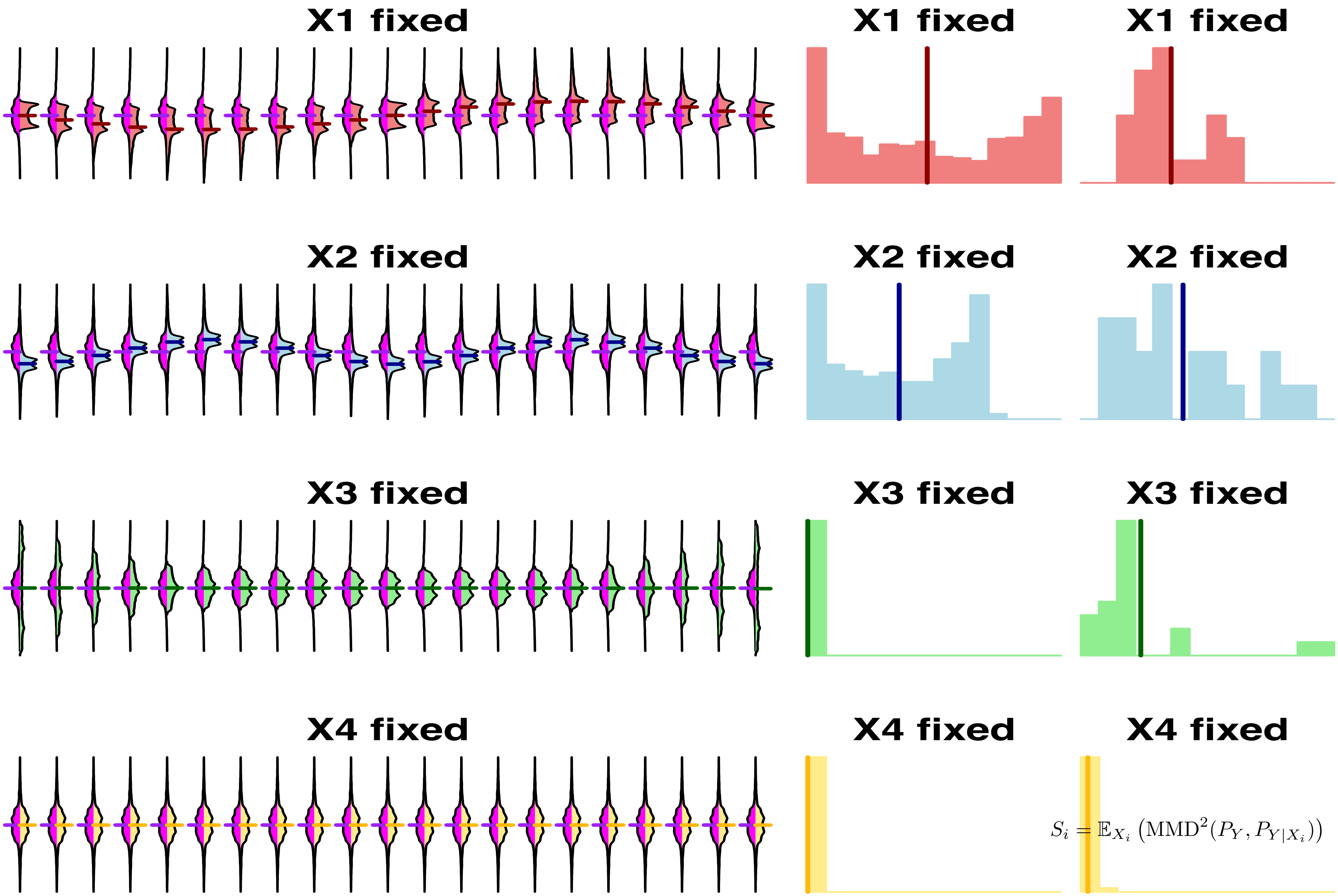
$$S_i = \mathbb{E}_{X_i}\left(d(P_Y, P_{Y|X_i})\right)$$

➔ If we use the MMD distance

$$d(P_Y, P_{Y|X_i}) = \mathrm{MMD}^2(P_Y, P_{Y|X_i})$$

$$\boxed{S_i = \mathbb{E}_{X_i}\left(\mathrm{MMD}^2(P_Y, P_{Y|X_i})\right)}$$

$$S_i = \int_\Omega k(y, y') \left[p_Y(y) - p_{Y|X_i=x_i}(y)\right] \left[p_Y(y') - p_{Y|X_i=x_i}(y')\right] p_{X_i}(x_i) dy dy' dx_i$$


X1 fixed
X1 fixed
X1 fixed
X2 fixed
X2 fixed
X2 fixed
X3 fixed
X3 fixed
X3 fixed
X4 fixed
X4 fixed
X4 fixed
$S_i = \mathbb{E}_{X_i}\left(\mathrm{MMD}^2(P_Y, P_{Y|X_i})\right)$

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

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

- But in practice you can choose any kernel, including

$$k(y, y') = \langle y, y' \rangle \overset{\mathrm{1D}}{=} yy'$$

Feature map is identity
Comparison through means only

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

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

- But in practice you can choose any kernel, including

$$k(y, y') = \langle y, y' \rangle \stackrel{\mathrm{1D}}{=} yy'$$

Feature map is identity
Comparison through means only

$$\mathbb{E}\left(\mathrm{MMD}^2(P_Y, P_{Y|X_i})\right) = \mathrm{Var}(\mathbb{E}(Y|X_i))$$

Unnormalized Sobol index

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

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

- But in practice you can choose any kernel, including

$$k(y, y') = \langle y, y' \rangle \overset{\mathrm{1D}}{=} yy'$$

Feature map is identity
Comparison through means only

- This is thus a natural extension of Sobol

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

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

- But in practice you can choose any kernel, including

$$k(y, y') = \langle y, y' \rangle \overset{\mathrm{1D}}{=} yy'$$

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**

$$\mathrm{Var}(Y) = \sum_{u \subseteq \{1,\ldots,d\}, u \neq \emptyset} g_u$$

$$g_u = \sum_{v \subseteq u} (-1)^{|u|-|v|} \mathrm{Var}\left(\mathbb{E}(Y|X_v)\right)$$

RKHS EMBEDDING: DECOMPOSITION I

➔ **Theorem (D. 2016)**

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

$$g_u = \sum_{v \subseteq u} (-1)^{|u|-|v|} \mathbb{E}\left(\mathrm{MMD}^2\left(P_{Y|X_v}, P_Y\right)\right)$$

RKHS EMBEDDING: DECOMPOSITION I

➔ Theorem (D. 2016)

$$\mathbb{E}\left(\mathrm{MMD}^2\left(P_{Y|X_{1:d}}, P_Y\right)\right) = \sum_{u \subseteq \{1,\dots,p\}, u \neq \emptyset} g_u$$

$$g_u = \sum_{v \subseteq u} (-1)^{|u|-|v|} \mathbb{E}\left(\mathrm{MMD}^2\left(P_{Y|X_v}, P_Y\right)\right)$$

➔ MMD sensitivity indices

$$\boxed{S_u^{\mathrm{MMD}} = \frac{\sum_{v \subseteq u} (-1)^{|u|-|v|} \mathbb{E}\left(\mathrm{MMD}^2\left(P_{Y|X_v}, P_Y\right)\right)}{\mathbb{E}\left(\mathrm{MMD}^2\left(P_{Y|X_{1:d}}, P_Y\right)\right)}}$$

RKHS EMBEDDING: DECOMPOSITION I

➔ 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, 1st order MMD indices are given by

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

Linear combination of the Sobol index for a (potential) infinity of transformations of the output (i.e. features)

RKHS EMBEDDING: ESTIMATION

➔ **Standard MMD estimation**

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

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

RKHS EMBEDDING: ESTIMATION

➔ Standard MMD estimation

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

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

➔ What about the MMD sensitivity index ?

$$S_i = \mathbb{E}_{X_i} \left(\mathrm{MMD}^2(P_Y, P_{Y|X_i}) \right)$$

- Brute-force Monte-Carlo very expensive
- Possible to use Pick & Freeze estimation
- Ongoing investigation of replicated designs to get rid of the input dimension

RKHS EMBEDDING: FEATURE SELECTION

➔ **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**

$$
\begin{aligned}
S_i^{KL} &= \int p_{Y|X_i=x}(y) \ln\left(\frac{p_{Y|X_i=x}(y)}{p_Y(y)}\right) p_{X_i}(x)dxdy \\
&= \int p_{Y,X_i}(y,x) \ln\left(\frac{p_{Y,X_i}(y,x)}{p_Y(y)p_{X_i}(x)}\right) dxdy = I(X_i;Y)
\end{aligned}
$$

Mutual Information

- **In this case, the sensitivity index is a dependence measure between random variables**

RKHS EMBEDDING: FEATURE SELECTION

- **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 ?

RKHS EMBEDDING: FEATURE SELECTION

- **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 ?

$$\begin{aligned}
\mathrm{MMD}^2\left(P_{Y,X}, P_Y P_X\right) &= \left(\sup_{f \in F}\left[\mathbb{E}_{P_{XY}} f(x,y) - \mathbb{E}_{P_X P_Y} f(x,y)\right]\right)^2 \\
&= \|\mu_{P_{XY}} - \mu_{P_X P_Y}\|^2_{\mathcal{F}\times\mathcal{G}} \\
&= \mathrm{HSIC}(X,Y)
\end{aligned}$$

Hilbert-Schmidt Independence Criterion
Gretton 2005

RKHS EMBEDDING: FEATURE SELECTION

➔ **HSIC estimation from a sample of the joint distribution**

$$\widehat{\mathrm{HSIC}}(X,Y) = \frac{1}{n^2}\mathrm{trace}\,(KHLH)$$

$$[K]_{ij} = k_{\mathcal{X}}(x_i, x_j) \qquad [L]_{ij} = k_{\mathcal{Y}}(y_i, y_j) \qquad [H]_{ij} = \delta_{ij} - \frac{1}{n}$$

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

➔ **Several feature selection techniques based on this measure**

- Song et al. (2007a,b,c), Balasubramanian et al. (2013), Yamada et al. 2013

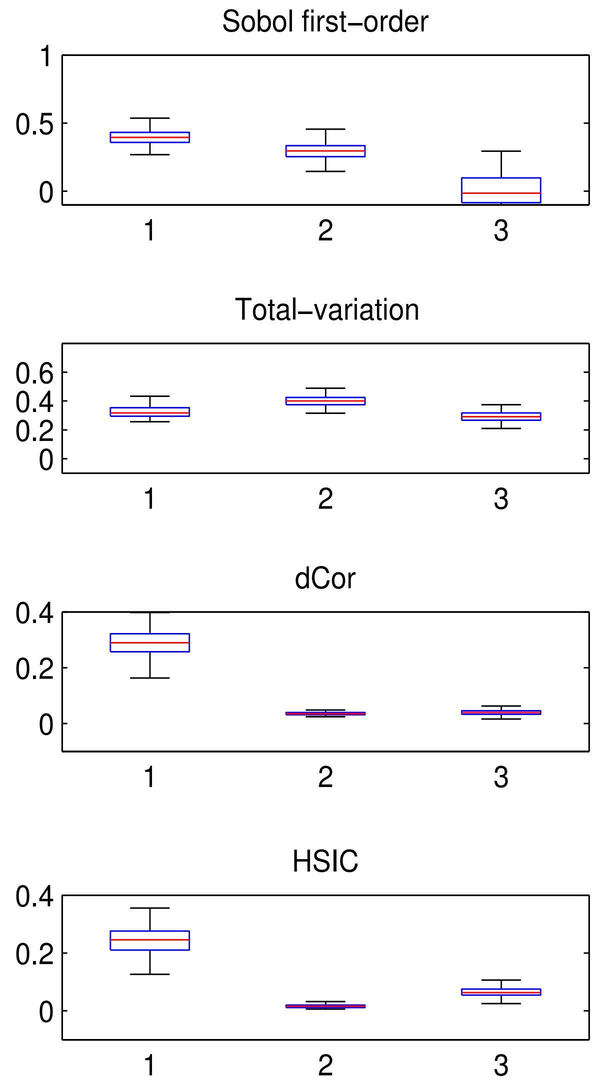
RKHS EMBEDDING: FEATURE SELECTION

➔ **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)



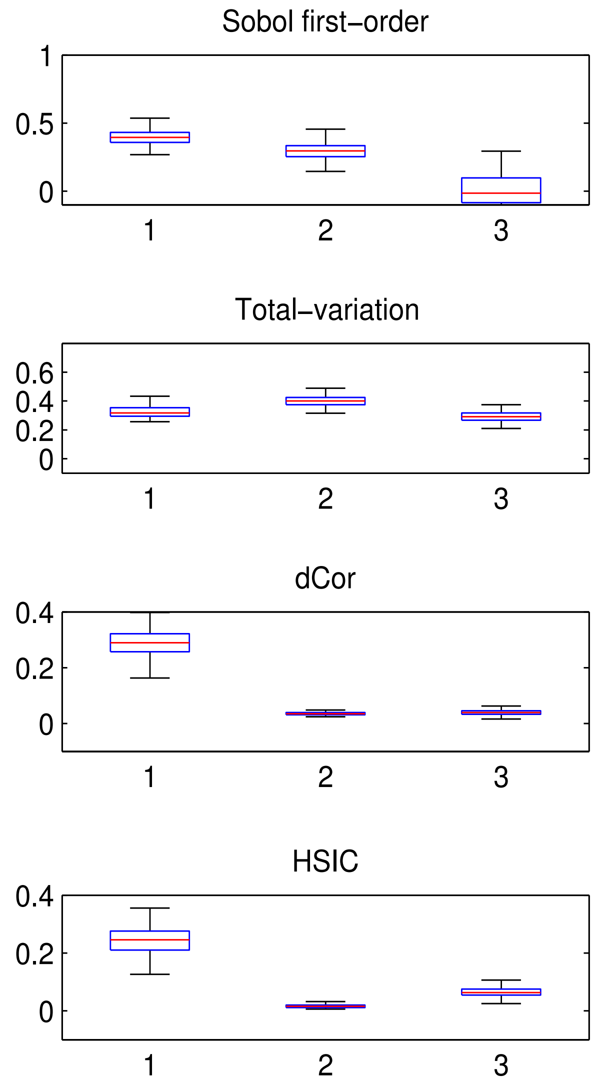
SAFRAN
AEROSPACE · DEFENCE · SECURITY

RKHS EMBEDDING: FEATURE SELECTION

- **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)

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

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

If the kernel on each input satisfies

$$\int_{\mathcal{X}} k_{\mathcal{X}}(x, x') dP_X(x) = 1$$

$$\begin{aligned}
k_{\mathcal{X}} &= 1 + k^0_{\mathcal{X}} \\
k^0_{\mathcal{X}}(x, x') &= k(x, x') - \frac{\int_{\mathcal{X}} k(x, x') dP_X(x') \int_{\mathcal{X}} k(x, x') dP_X(x)}{\iint_{\mathcal{X}\times\mathcal{X}} k(x, x') dP_X(x) dP_X(x')} \\
k^0_{\mathcal{X}}(x, x') &= k(x, x') - \int_{\mathcal{X}} k(x, x') dP_X(x') - \int_{\mathcal{X}} k(x, x') dP_X(x) \\
&\quad + \iint_{\mathcal{X}\times\mathcal{X}} k(x, x') dP_X(x) dP_X(x')
\end{aligned}$$

RKHS EMBEDDING: DECOMPOSITION II

➔ **Theorem (D. 2016)**

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

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

$$\boxed{S_u^{\mathrm{HSIC}} = \frac{\sum_{v \subseteq u} (-1)^{|u|-|v|} \mathrm{HSIC}\left(Y, X_v\right)}{\mathrm{HSIC}\left(Y, X_{1:d}\right)}}$$

RKHS EMBEDDING: DECOMPOSITION II

➔ 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)

RKHS EMBEDDING: DECOMPOSITION II

➔ 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

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

- 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**

Uniform inputs

$$k_{\mathcal{X}}(x, x') \to \delta(x, x')$$

ex: $k_{\mathcal{X}}(x, x') = \frac{1}{\sqrt{2\pi a^2}} \exp\left(\frac{1}{2a^2}(x - x')^2\right), \quad a \to 0$

$$\boxed{S_u^{\mathrm{HSIC}} \longrightarrow S_u^{\mathrm{MMD}}}$$

RKHS EMBEDDING: DECOMPOSITION II

➔ 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**

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

Uniform inputs $$k_{\mathcal{X}}(x,x') \rightarrow \delta(x,x')$$

$$\text{ex: } k_{\mathcal{X}}(x,x') = \frac{1}{\sqrt{2\pi a^2}} \exp\left(\frac{1}{2a^2}(x-x')^2\right), \quad a \rightarrow 0$$

$$\boxed{S_u^{\mathrm{HSIC}} \longrightarrow S_u^{\mathrm{MMD}}}$$

$$k_{\mathcal{Y}}(y,y') = yy' \qquad \boxed{S_u^{\mathrm{MMD}} = S_u^{\mathrm{Sobol}}}$$

RKHS EMBEDDING: LET'S PLAY WITH KERNELS

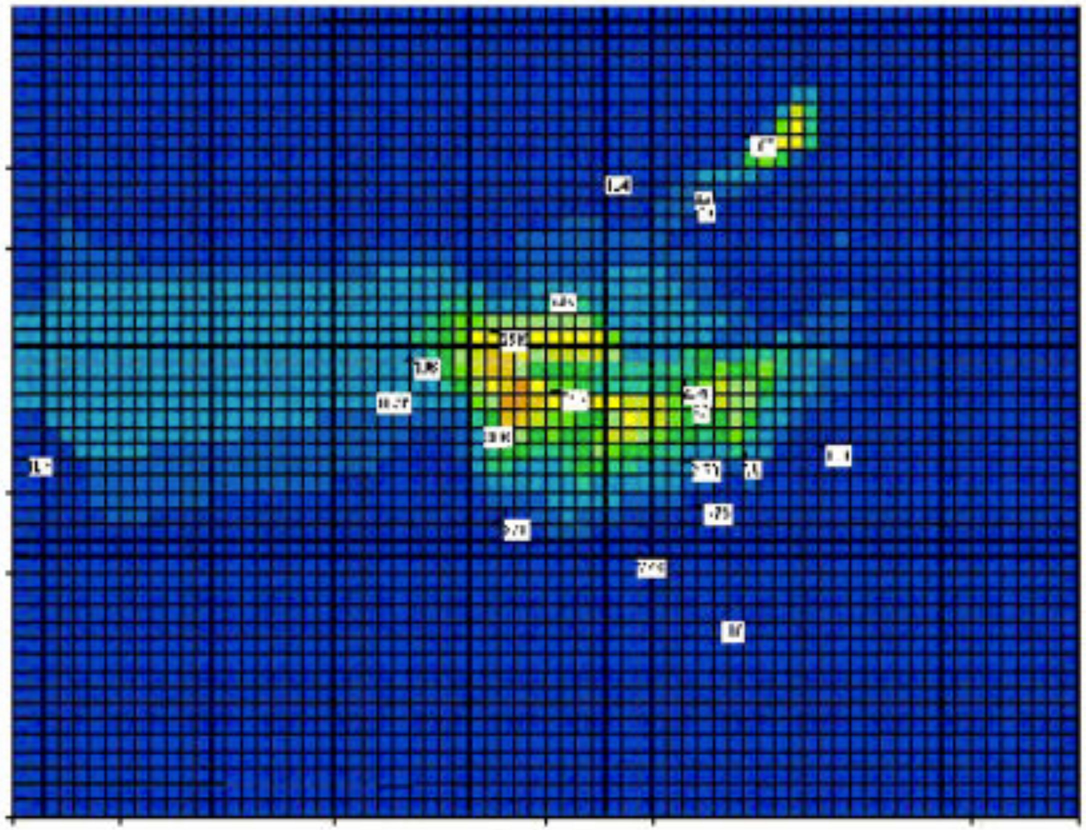
- **The RKHS point of view comes with a huge literature and dedicated kernels**
  - If your inputs or outputs are vectors, curves, texts, images, timeseries, DNA sequences, probability distributions, … there is a kernel available
    - **We then have a generic GSA framework which can handle them, with a decomposition into main effects and interactions**

RKHS EMBEDDING: LET'S PLAY WITH KERNELS

➔ **Example 1: migration of strontium 90 in a storage site (Marthe testcase CEA)**

- Inputs
  - 20 geological parameters

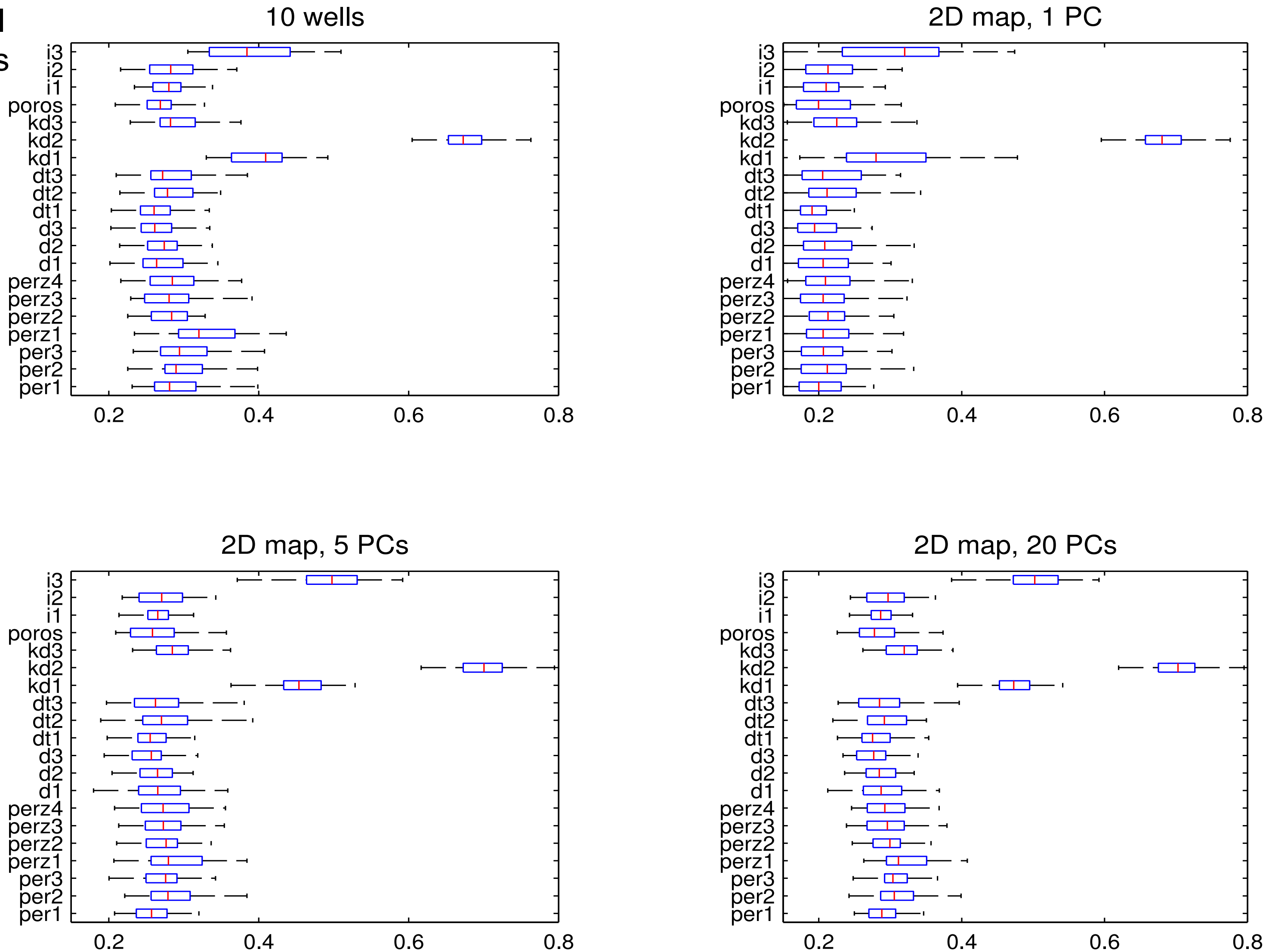
- Outputs
  - Strontium concentration at 10 observation wells
  - 2D maps of concentration (64x64=4096 pixels)



Marrel et al. 2011

RKHS EMBEDDING: LET'S PLAY WITH KERNELS

Gaussian kernel in 10 dimensions (wells)



PCA kernel on 2D maps

D. 2014

SAFRAN AEROSPACE · DEFENCE · SECURITY

RKHS EMBEDDING: LET'S PLAY WITH KERNELS

➔ Example 2: Optimization

$$\min_{x \in \mathcal{X}} f(x)$$

$$\text{s.t.} \quad h(x) \leq 0$$

- 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

RKHS EMBEDDING: LET'S PLAY WITH KERNELS

➔ Example 2: Optimization

$$\min_{x \in \mathcal{X}} f(x)$$

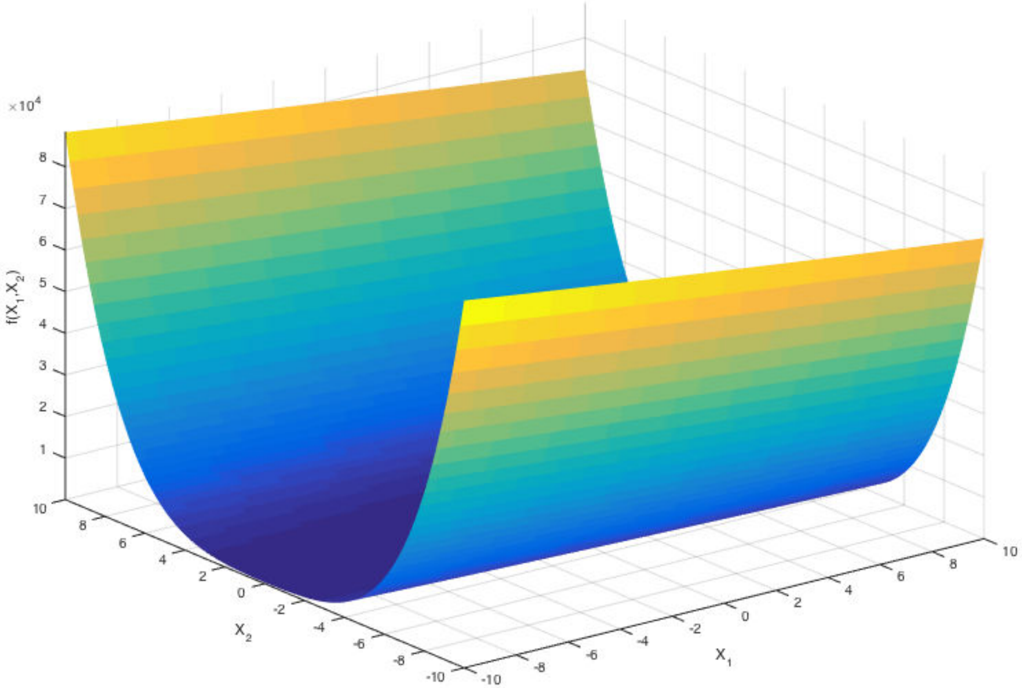
$$\text{s.t.} \quad h(x) \leq 0$$

- 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**

RKHS EMBEDDING: LET'S PLAY WITH KERNELS

➔ Example 2: Optimization



Dixon-Price function

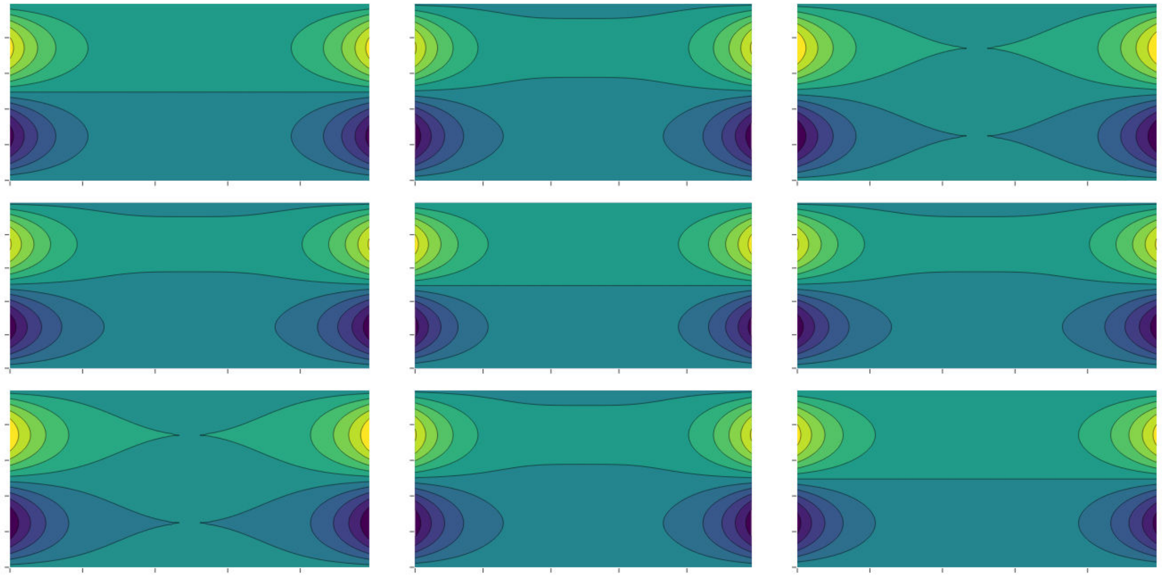
One variable has a sensitivity index equal to 1

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

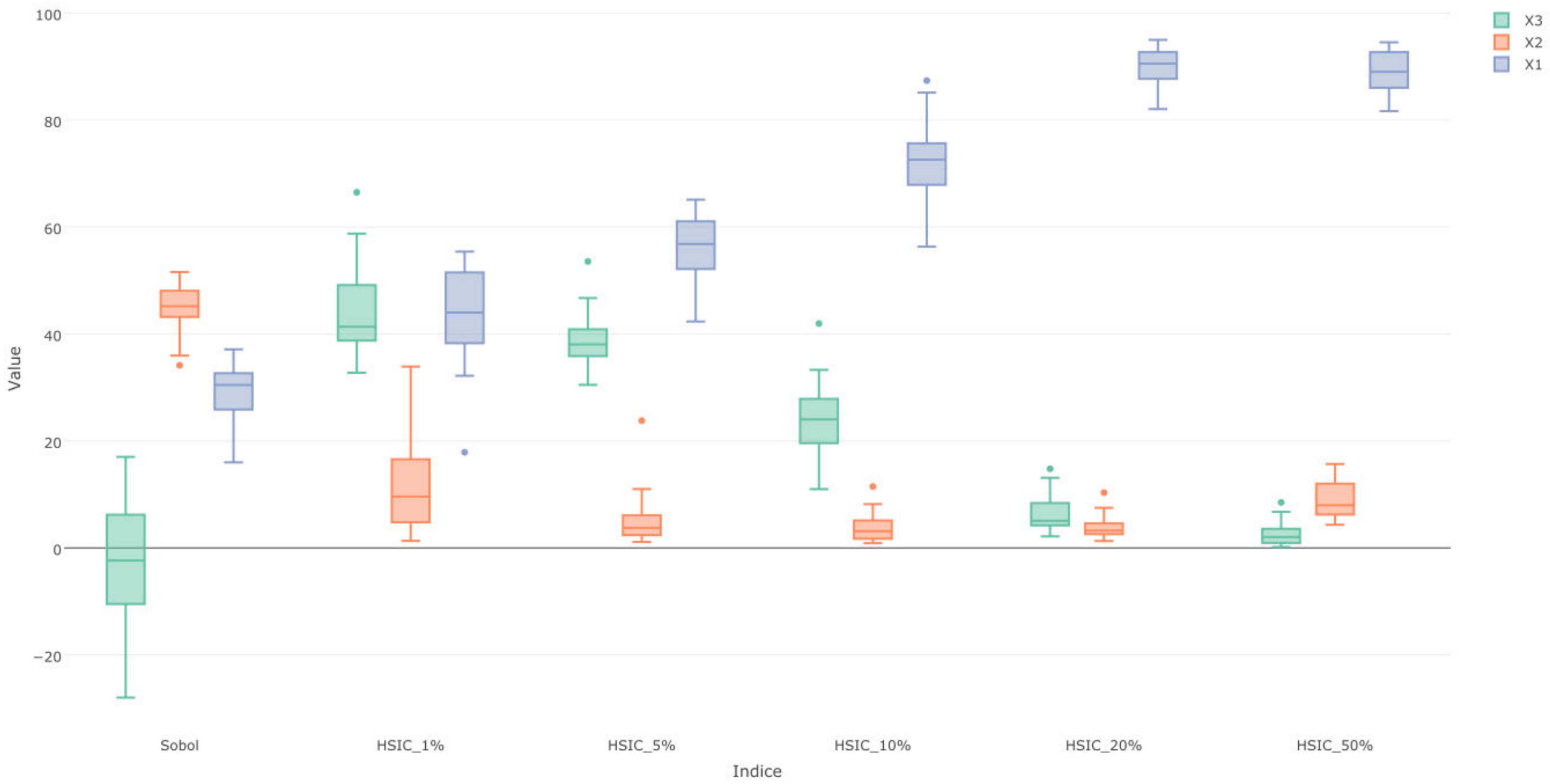
Variable 2 has a significant Sobol index

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

RKHS EMBEDDING: LET'S PLAY WITH KERNELS

➔ Example 2: Optimization & dedicated kernels



Master's internship of Adrien Spagnol @ IRT SystemX

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

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

$$k_{\mathcal{X}}(x, x') \to \delta(x, x')$$

ex: $k_{\mathcal{X}}(x, x') = \frac{1}{\sqrt{2\pi a^2}} \exp\left(\frac{1}{2a^2}(x-x')^2\right), \quad a \to 0$

$$\boxed{S_u^{\mathrm{HSIC}} \longrightarrow S_u^{\mathrm{MMD}}}$$

$$k_{\mathcal{Y}}(y, y') = yy' \qquad \boxed{S_u^{\mathrm{MMD}} = S_u^{\mathrm{Sobol}}}$$

SAFRAN AEROSPACE · DEFENCE · SECURITY

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

$$k_{\mathcal{X}_{-i}}(x, x') \to \delta(x, x') \qquad k_{\mathcal{X}_i}(x, x') \to \delta'(x, x')$$

$$k_{\mathcal{Y}}(y, y') = yy'$$

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

$$k_{\mathcal{X}_{-i}}(x, x') \to \delta(x, x') \qquad k_{\mathcal{X}_i}(x, x') \to \delta'(x, x')$$

$$k_{\mathcal{Y}}(y, y') = yy'$$

$$\boxed{S_i^{\mathrm{HSIC}} \to \int_\Omega \left(\frac{\partial \eta(x)}{\partial x_i} \right)^2 dx}$$

We recover the 1^{st} order DGSM indices !

SAFRAN AEROSPACE · DEFENCE · SECURITY

CONCLUSION

➔ **New sensitivity index which generalizes GSA through the use of kernels**

$$S_u^{\mathrm{HSIC}} = \frac{\sum_{v \subseteq u} (-1)^{|u|-|v|} \mathrm{HSIC}\left(Y, X_v\right)}{\mathrm{HSIC}\left(Y, X_{1:d}\right)}$$

- 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

CONCLUSION

➔ **I honestly think there is potentiel 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

Groupement de Recherche
MASCOT-NUM
Méthodes d'Analyse Stochastique pour les COdes et Traitements NUMériques

COME SEE US AT THE NEXT GDR ANNUAL MEETING

SAFRAN – MASSY
22/03 – 24/03 2017

SAFRAN AEROSPACE · DEFENCE · SECURITY

REFERENCES

Balasubramanian, K., Sriperumbudur, B., & Lebanon, G. (2013). Ultrahigh dimensional feature screening via rkhs embeddings. In Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (pp. 126-134).

Baucells, M., & Borgonovo, E. (2013). Invariant probabilistic sensitivity analysis.Management Science, 59(11), 2536-2549.

Borgonovo, E. (2007). A new uncertainty importance measure. Reliability Engineering & System Safety, 92(6), 771-784.

Da Veiga, S. (2015). Global sensitivity analysis with dependence measures.Journal of Statistical Computation and Simulation, 85(7), 1283-1305.

Durrande, N., Ginsbourger, D., Roustant, O., & Carraro, L. (2013). ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis.Journal of Multivariate Analysis, 115, 57-67.

Fort, J. C., Klein, T., & Rachdi, N. (2015). New sensitivity analysis subordinated to a contrast. To appear in Communications in Statistics-Theory and Methods.

Fukumizu, K., Gretton, A., Sun, X., & Schölkopf, B. (2007, December). Kernel Measures of Conditional Dependence. In NIPS (Vol. 20, pp. 489-496).

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., & Schölkopf, B. (2005). Kernel methods for measuring independence. The Journal of Machine Learning Research, 6, 2075-2129.

Gretton, A. (2012). Kernel Distribution Embeddings: Theory and Applications, Slides from a talk at the Oxford statistics department, http://www.gatsby.ucl.ac.uk/~gretton/papers/oxford12.pdf

Iooss, B., & Lemaître, P. (2015). A review on global sensitivity analysis methods. In Uncertainty Management in Simulation-Optimization of Complex Systems (pp. 101-122). Springer US.

Jiao, Y., & Vert, J. P. (2015). The Kendall and Mallows kernels for permutations. In Proceedings of the 32nd International Conference on Machine Learning (ICML-15) (pp. 1935-1944)

Krzykacz-Hausmann, B. (2001). Epistemic sensitivity analysis based on the concept of entropy. In International symposium on sensitivity analysis of model output (pp. 53-57).

REFERENCES

Lamboni, M., Iooss, B., Popelin, A. L., & Gamboa, F. (2013). Derivative-based global sensitivity measures: general links with Sobol'indices and numerical tests. Mathematics and Computers in Simulation, 87, 45-54.

Lemaître, P., Sergienko, E., Arnaud, A., Bousquet, N., Gamboa, F., & Iooss, B. (2015). Density modification-based reliability sensitivity analysis. Journal of Statistical Computation and Simulation, 85(6), 1200-1223.

Marrel, A., Iooss, B., Jullien, M., Laurent, B., & Volkova, E. (2011). Global sensitivity analysis for models with spatially dependent outputs. Environmetrics,22(3), 383-397.

Smola, A., Gretton, A., Song, L., & Schölkopf, B. (2007, October). A Hilbert space embedding for distributions. In Algorithmic learning theory (pp. 13-31). Springer Berlin Heidelberg.

Song, L., Smola, A., Gretton, A., Borgwardt, K. M., & Bedo, J. (2007, June). Supervised feature selection via dependence estimation. In Proceedings of the 24th international conference on Machine learning (pp. 823-830). ACM.

Song, L., Smola, A., Gretton, A., & Borgwardt, K. M. (2007, June). A dependence maximization view of clustering. In Proceedings of the 24th international conference on Machine learning (pp. 815-822). ACM.

Song, L., Bedo, J., Borgwardt, K. M., Gretton, A., & Smola, A. (2007). Gene selection via the BAHSIC family of algorithms. Bioinformatics, 23(13), i490-i498.

Song, L. (2008). Learning via Hilbert space embedding of distributions. PhD Thesis, University of Sidney.

Song, L., Huang, J., Smola, A., & Fukumizu, K. (2009, June). Hilbert space embeddings of conditional distributions with applications to dynamical systems. In Proceedings of the 26th Annual International Conference on Machine Learning (pp. 961-968). ACM.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Lanckriet, G. R., & Schölkopf, B. (2008, July). Injective Hilbert Space Embeddings of Probability Measures. InCOLT (Vol. 21, pp. 111-122).

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., & Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. The Journal of Machine Learning Research, 11, 1517-1561.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., & Sugiyama, M. (2014). High-dimensional feature selection by feature-wise kernelized lasso. Neural computation, 26(1), 185-207.