Indices HSIC pour l'analyse de sensibilité

Introduction & avancées récentes

Sébastien Da Veiga — ENSAI / CREST — Journées Mexico 2022

Outline

- Context – Global Sensitivity Analysis (GSA)
- Generalized GSA via kernel embedding of probability distributions
- Conclusion & outlook

CONTEXT

GLOBAL SENSITIVITY ANALYSIS

Sensitivity analysis: Sobol' indices arise from a functional ANOVA decomposition

Theorem 1 (ANOVA decomposition (Hoeffding, 1948; Antoniadis, 1984)). *Assume that* $\eta : \mathcal{X}_1 \times \ldots \times \mathcal{X}_d \to \mathcal{Y}$ *is a square integrable function of* d *independent random variables* $X_1, \ldots, X_d$. *Then* η *admits a decomposition*

$$Y = \eta(X_1, \ldots, X_d) = \sum_{A \subseteq \mathcal{P}_d} \eta_A(\mathbf{X}_A),$$

with η_A *depending only on the variables* $\mathbf{X}_A$ *and satisfying*

(a) $\eta_\emptyset = \mathbb{E}(Y)$,

(b) $\mathbb{E}_{X_l}(\eta_A(\mathbf{X}_A)) = 0$ *if* $l \in A$,

(c) $\eta_A(\mathbf{X}_A) = \sum_{B \subset A} (-1)^{|A|-|B|} \mathbb{E}(Y|\mathbf{X}_B)$.

Furthermore, (b) implies that all the terms η_A *in the decomposition are mutually orthogonal. As a consequence, the output variance can be decomposed as*

$$\operatorname{Var} Y = \sum_{A \subseteq \mathcal{P}_d} \operatorname{Var} \eta_A(\mathbf{X}_A) = \sum_{A \subseteq \mathcal{P}_d} V_A \tag{1}$$

where

$$V_A = \sum_{B \subset A} (-1)^{|A|-|B|} \operatorname{Var} \mathbb{E}(Y|\mathbf{X}_B). \tag{2}$$

Sensitivity analysis: Sobol' indices arise from a functional ANOVA decomposition

Definition 1 (Sobol' indices (Sobol', 1993)). *Under the same assumptions of Theorem 1, the Sobol' sensitivity index associated to a subset A of input variables is defined as*

$$S_A = \frac{V_A}{\operatorname{Var} Y}, \tag{3}$$

A is a subset of input variables

while the total Sobol' index associated to A is

$$S_A^T = \sum_{B \subseteq \mathcal{P}_d,\, B \cap A \neq \emptyset} S_B. \tag{4}$$

In particular, the first-order Sobol' index of an input X_l writes

$$S_l = \frac{\operatorname{Var} \mathbb{E}(Y | X_l)}{\operatorname{Var} Y}$$

Impact of an input alone

and its total Sobol' index is given by

$$S_l^T = \sum_{B \subseteq \mathcal{P}_d,\, l \in B} S_B = 1 - \frac{\operatorname{Var} \mathbb{E}(Y | \mathbf{X}_{-l})}{\operatorname{Var} Y}.$$

Impact of an input through all its potential interactions with others

Finally, the ANOVA decomposition (1) readily provides an interpretation of Sobol' indices as a percentage of explained output variance, i.e.

$$\sum_{A \subseteq \mathcal{P}_d} S_A = 1. \tag{5}$$

Interpretation as percentage

Sensitivity analysis: Sobol' indices

- **Sobol' indices**
  - ➡ The impact of each input can be quantitatively assessed
    - First-order effect
    - Total effect including also all possible interactions with other inputs
    - Pure interactions can be properly defined

$$S_{ll'} = \frac{\mathrm{Var}\,\mathbb{E}(Y|X_l,\, X_{l'}) - \mathrm{Var}\,\mathbb{E}(Y|X_l) - \mathrm{Var}\,\mathbb{E}(Y|X_{l'})}{\mathrm{Var}\,Y} = \frac{\mathrm{Var}\,\mathbb{E}(Y|X_l,\, X_{l'})}{\mathrm{Var}\,Y} - S_l - S_{l'}$$

First-order effects can be propertly subtracted

Sensitivity analysis: Sobol' indices

- **Sobol' indices**

Limitations

Assumption of independent inputs (more on this at the end)

Impact on output variance only

Outputs may not be scalars

Cannot be used for screening due to computational cost

Sensitivity analysis: Sobol' indices

- **Sobol' indices**

Limitations

Assumption of independent inputs (more on this at the end) ➡ **We will talk about Shapley later**

Impact on output variance only

Outputs may not be scalars

Cannot be used for screening due to computational cost

Sensitivity analysis: Sobol' indices

- **Sobol' indices**

Limitations

Assumption of independent inputs (more on this at the end) ➡ We will talk about Shapley later

Impact on output variance only

Outputs may not be scalars ➡ **Moment-independent indices with kernels…**

Cannot be used for screening due to computational cost

Sensitivity analysis: Sobol' indices

- **Sobol' indices**

Limitations

Assumption of independent inputs (more on this at the end) ➡ We will talk about Shapley later

Impact on output variance only

Outputs may not be scalars

➡ Moment-independent indices with kernels…

Cannot be used for screening due to computational cost ➡ **… In particular, HSIC**

Sensitivity analysis: our journey today

	Independent inputs		
	Sobol		Moment-independent
	1st order	Total order	Density-based
Beyond variance	✗	✗	✓
ANOVA (ranking)	✓	✓	✗
Screening	✗	✓	✗
Estimation (given data + small data)	✓	✗	✗
Can handle dependent inputs	✗	✗	✗
Can handle any output type	✗	✗	✗

Sensitivity analysis: our journey today

	Independent inputs		
	Sobol		Moment-independent
	1st order	Total order	Density-based
Beyond variance	✗	✗	✓
ANOVA (ranking)	✓	✓	✗
Screening	✗	✓	✗
Estimation (given data + small data)	✓	✗	✗
Can handle dependent inputs	✗	✗	✗
Can handle any output type	✗	✗	✗

Useful for in-depth analysis, definition of interactions …

Sensitivity analysis: our journey today

	Independent inputs		
	Sobol		Moment-independent
	1st order	Total order	Density-based
Beyond variance	✗	✗	✓
ANOVA (ranking)	✓	✓	✗
Screening	✗	✓	✗
Estimation (given data + small data)	✓	✗	✗
Can handle dependent inputs	✗	✗	✗
Can handle any output type	✗	✗	✗

Useful for in-depth analysis, definition of interactions …

Both are necessary for practical screening

Sensitivity analysis: our journey today

	Independent inputs			
	Sobol		Moment-independent	
	1st order	Total order	Density-based	
Beyond variance	✗	✗	✓	Interesting for generalization
ANOVA (ranking)	✓	✓	✗	Useful for in-depth analysis, definition of interactions …
Screening	✗	✓	✗	Both are necessary for practical screening
Estimation (given data + small data)	✓	✗	✗	
Can handle dependent inputs	✗	✗	✗	
Can handle any output type	✗	✗	✗	Interesting for generalization

Sensitivity analysis: our journey today

	Independent inputs		
	Sobol		**Moment-independent**
	1st order	Total order	Density-based
Beyond variance	✗	✗	✓
ANOVA (ranking)	✓	✓	✗
Screening	✗	✓	✗
Estimation (given data + small data)	✓	✗	✗
Can handle dependent inputs	✗	✗	✗
Can handle any output type	✗	✗	✗

Today we will introduce several **new sensitivity indices** based on **kernels** which aim at improving this picture!

Sensitivity analysis: our journey today

	Independent inputs							
	Sobol		Moment-independent					
	1st order	Total order	Density-based	1st order MMD	Total order MMD	HSIC	1st order HSIC ANOVA	Total order HSIC ANOVA
Beyond variance	✗	✗	✓	✓	✓	✓	✓	✓
ANOVA (ranking)	✓	✓	✗	✓	✓	✗	✓	✓
Screening	✗	✓	✗	✗	✓	✓	✓	✓
Estimation (given data + small data)	✓	✗	✗	✓	✗	✓	✓	✓
Can handle dependent inputs	✗	✗	✗	✗	✗	✓	✗	✗
Can handle any output type	✗	✗	✗	✓	✓	✓	✓	✓

Kernel-based sensitivity analysis

Sensitivity analysis: other indices

- **Going beyond the variance 1: goal-oriented sensitivity analysis**
  - Indices based on contrast functions (Fort et al. 2014), in particular quantile-oriented indices
  - Reliability-based indices
  - Many industrial applications

- **Going beyond the variance 2: moment-independent indices**
  - Principle: Quantify the impact of an input parameter on the probability distribution of the output

$$\mathcal{S}_l^{TV} = \int |p_Y(y) - p_{Y|X_l=x}(y)| p_{X_l}(x) dx dy$$

Borgonovo 2007

$$\mathcal{S}_l^{KL} = \int p_{Y|X_l=x}(y) \ln\left(\frac{p_{Y|X_l=x}(y)}{p_Y(y)}\right) p_{X_l}(x) dx dy$$

Kraskov et al. 2001

Sensitivity analysis: general point of view

- **General framework for moment independent indices**

$$\mathcal{S}_l = \mathbb{E}_{X_l}\left(d(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right)$$

Baucells & Borgonovo 2013
D. 2015

  - If the output probability distribution and the conditional one are « close », the input parameter has little influence
  - Example: f-divergence (D. 2015, Rahman 2016), with particular cases TV & KL

Sensitivity analysis: general point of view

- **General framework for moment independent indices**

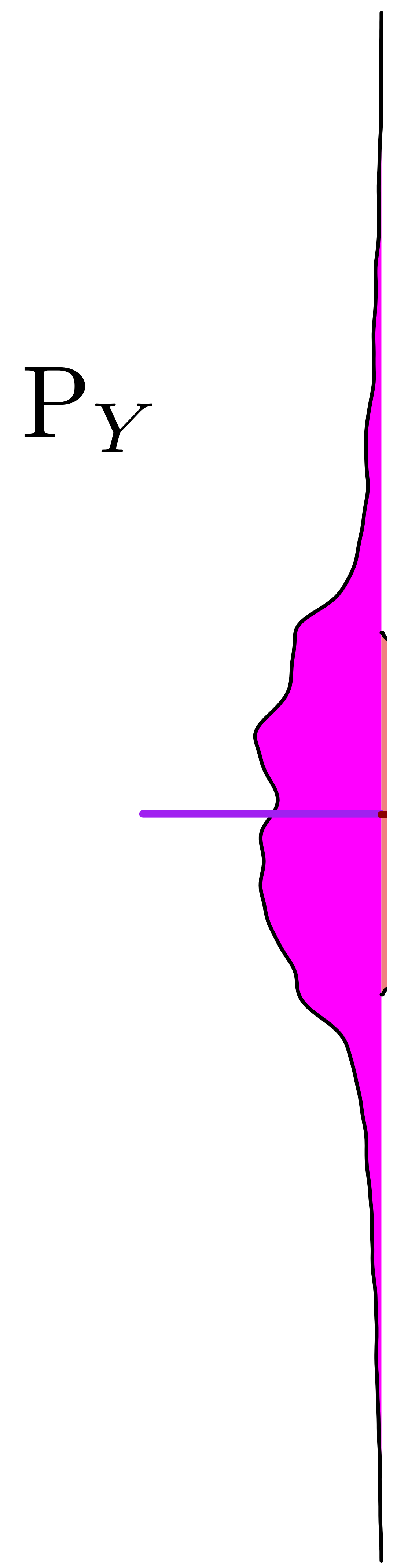
$$\mathcal{S}_l = \mathbb{E}_{X_l}\left(d(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right)$$

Baucells & Borgonovo 2013
D. 2015

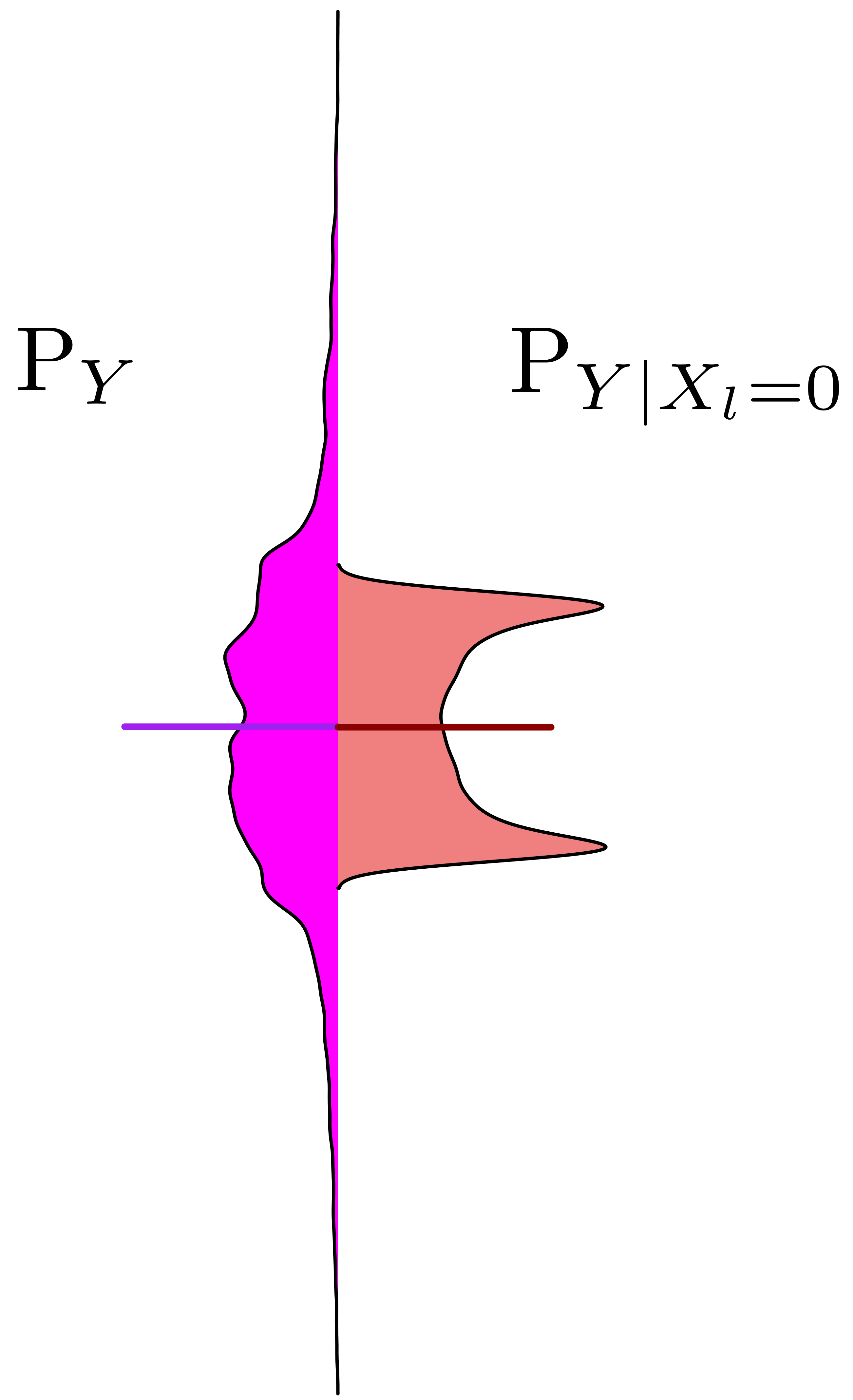
  - If the output probability distribution and the conditional one are « close », the input parameter has little influence
  - Example: f-divergence (D. 2015, Rahman 2016), with particular cases TV & KL
  - Toy example

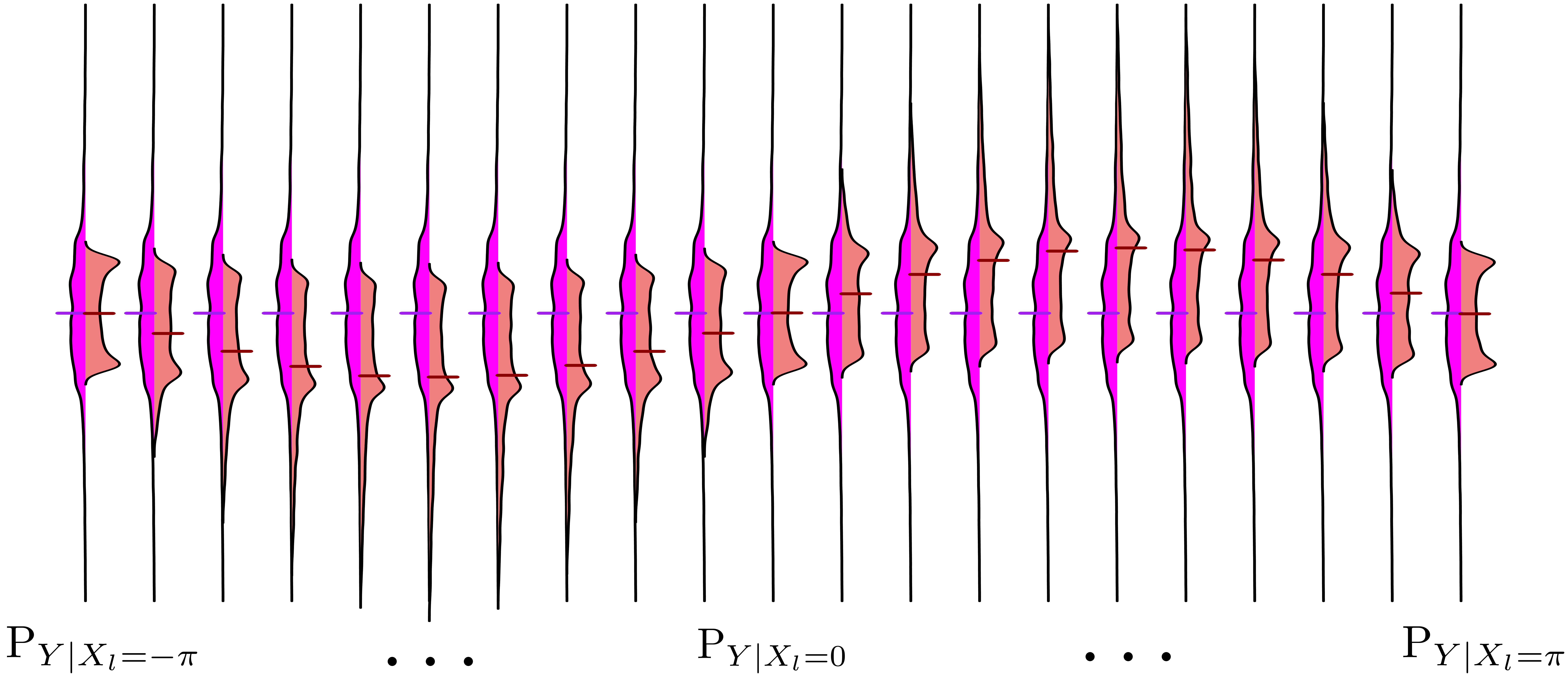
$$Y = \sin(X_1) + 7\sin(X_2)^2 + X_3^4 \sin(X_1)$$

$$X_l \sim \mathcal{U}(-\pi, \pi) \text{ for } l = 1, \dots, 4$$

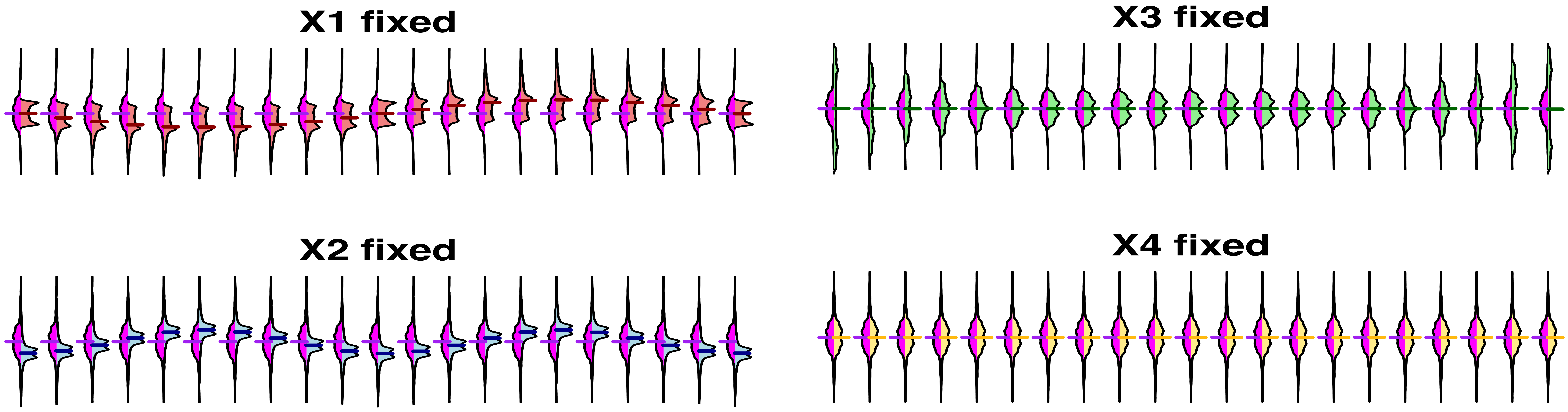


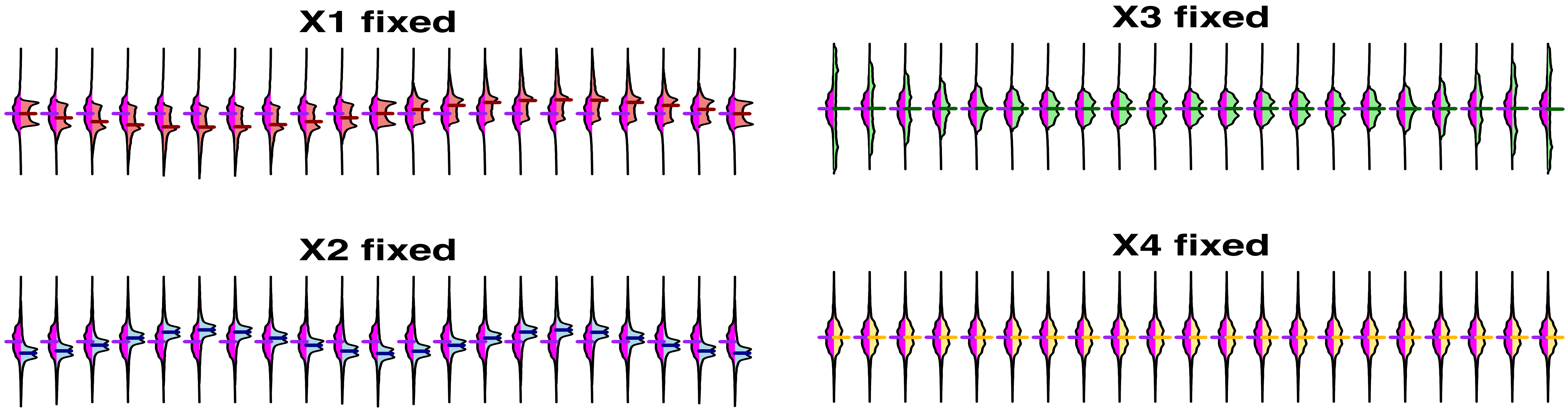
P_Y


P_Y
$P_{Y|X_l=0}$



$P_{Y|X_l=-\pi}$
...
$P_{Y|X_l=0}$
...
$P_{Y|X_l=\pi}$


X1 fixed
X3 fixed
X2 fixed
X4 fixed



Moment independent indices

➡ Pros

- They account for the whole effect of a parameter on the output distribution
- Density-based (many methods & packages)

➡ Cons

- Higher-order indices or outputs implies curse of dimensionality
- No ANOVA (« natural » normalization constant? Separation between interactions & main effects?)

$$\mathcal{S}_{ll'}^{TV} = \int |p_Y(y)p_{X_l}(x)p_{X_{l'}}(x') - p_{X_l,X_{l'},Y}(x,x',y)|dxdx'dy - \mathcal{S}_l^{TV} - \mathcal{S}_{l'}^{TV}$$

Does this make sense?

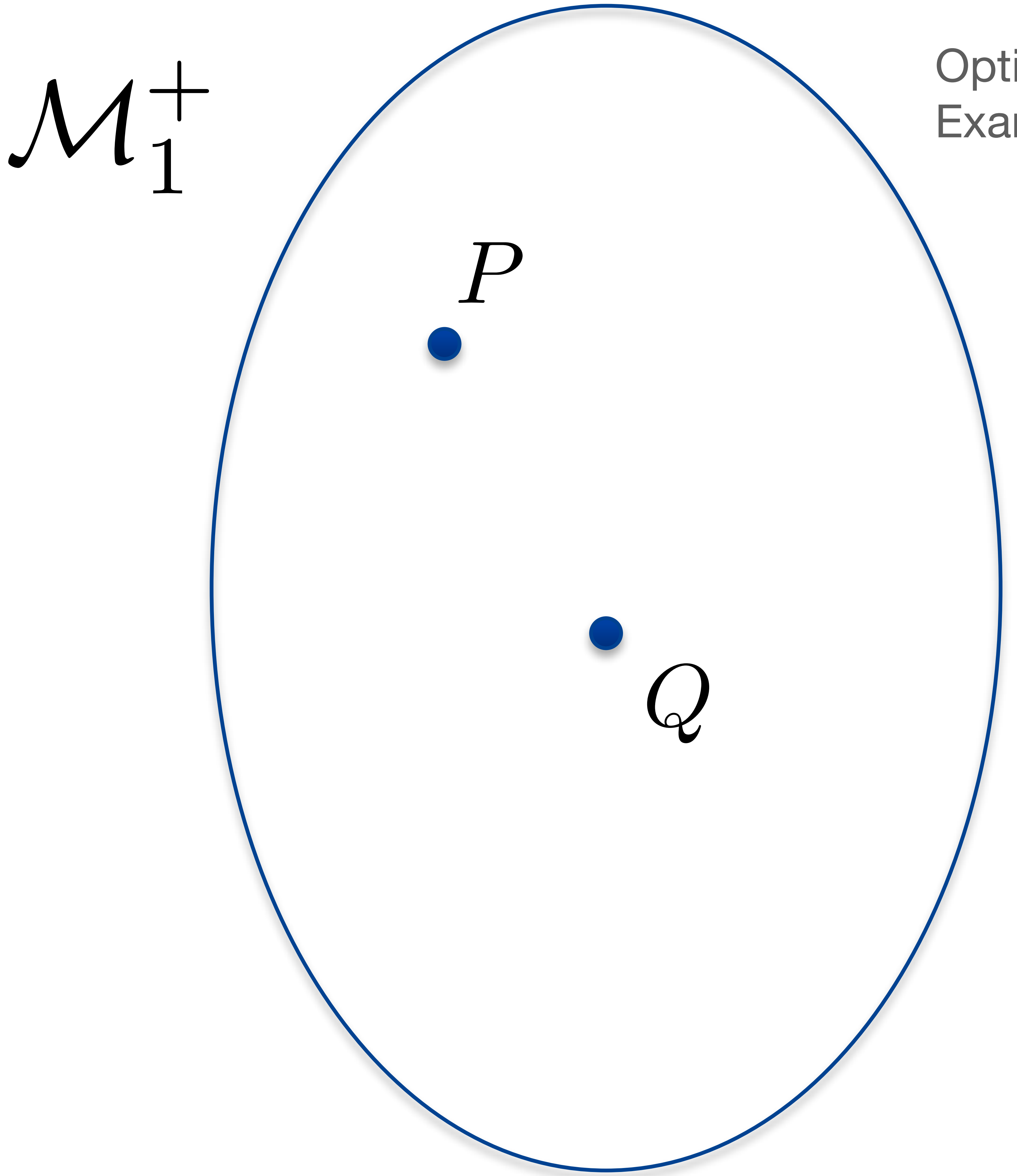
Sensitivity analysis: our journey today

- **Step 1: another look at moment-independent indices**
  - We will use a promising candidate for the distance
  - Theory of kernel-embedding of probability distributions
  - A new sensitivity index with ANOVA decomposition: MMD indices
- **Step 2: going further for screening**
  - We will introduce another kernel-based index, with much less computation cost: HSIC indices
  - With a recent powerful result = ANOVA decomposition also!
- **Step 3: handling dependence**
  - HSIC indices can be used, but without quantitative ranking
  - We propose kernel-based extension of Shapley effects

KERNEL-EMBEDDING OF PROBABILITY DISTRIBUTIONS

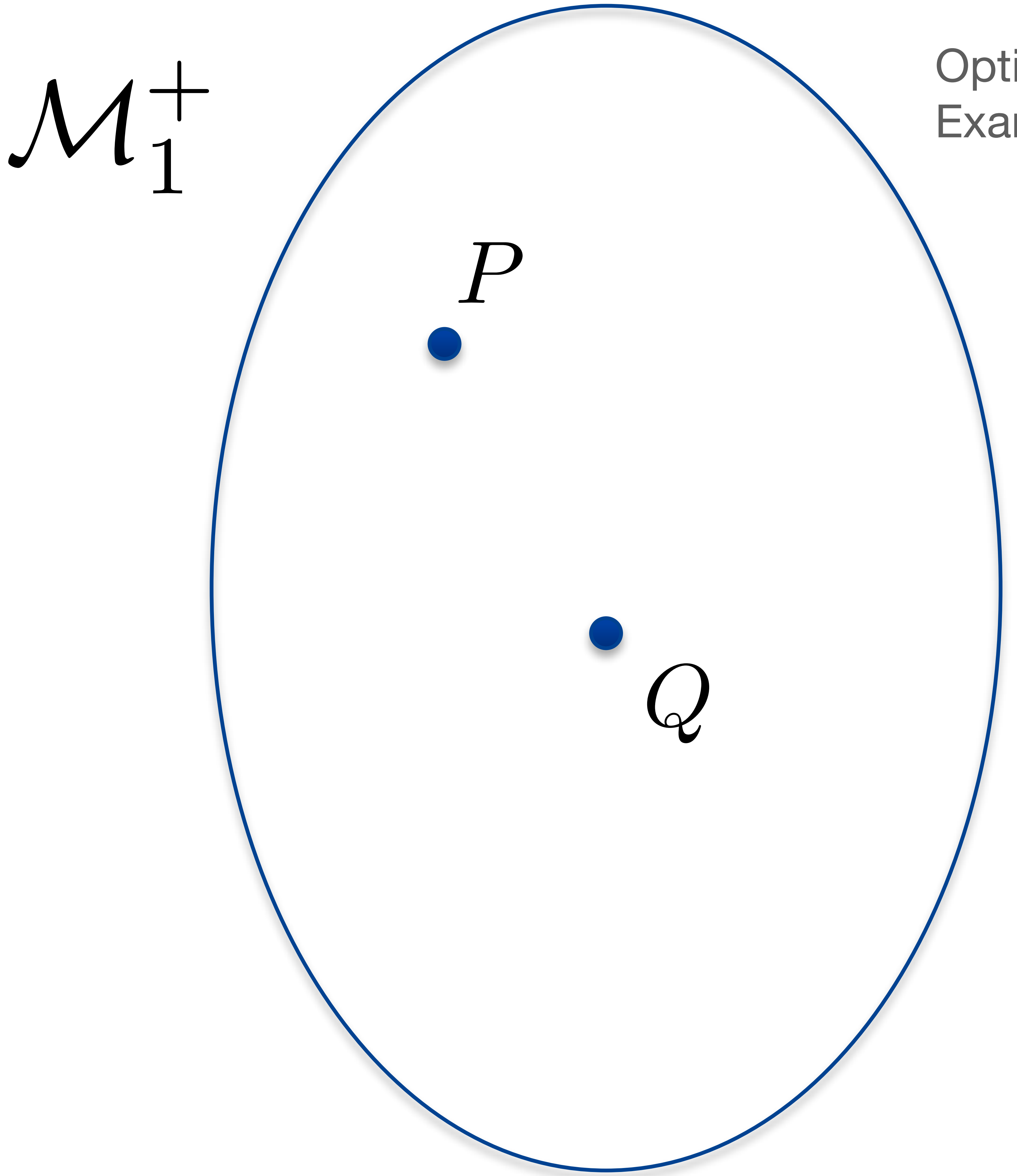
A VERY QUICK SUMMARY

Kernel-embedding of probability distributions



Option 1: work directly in the space of probability measures
Examples: KS, TV, KL, Hellinger, …

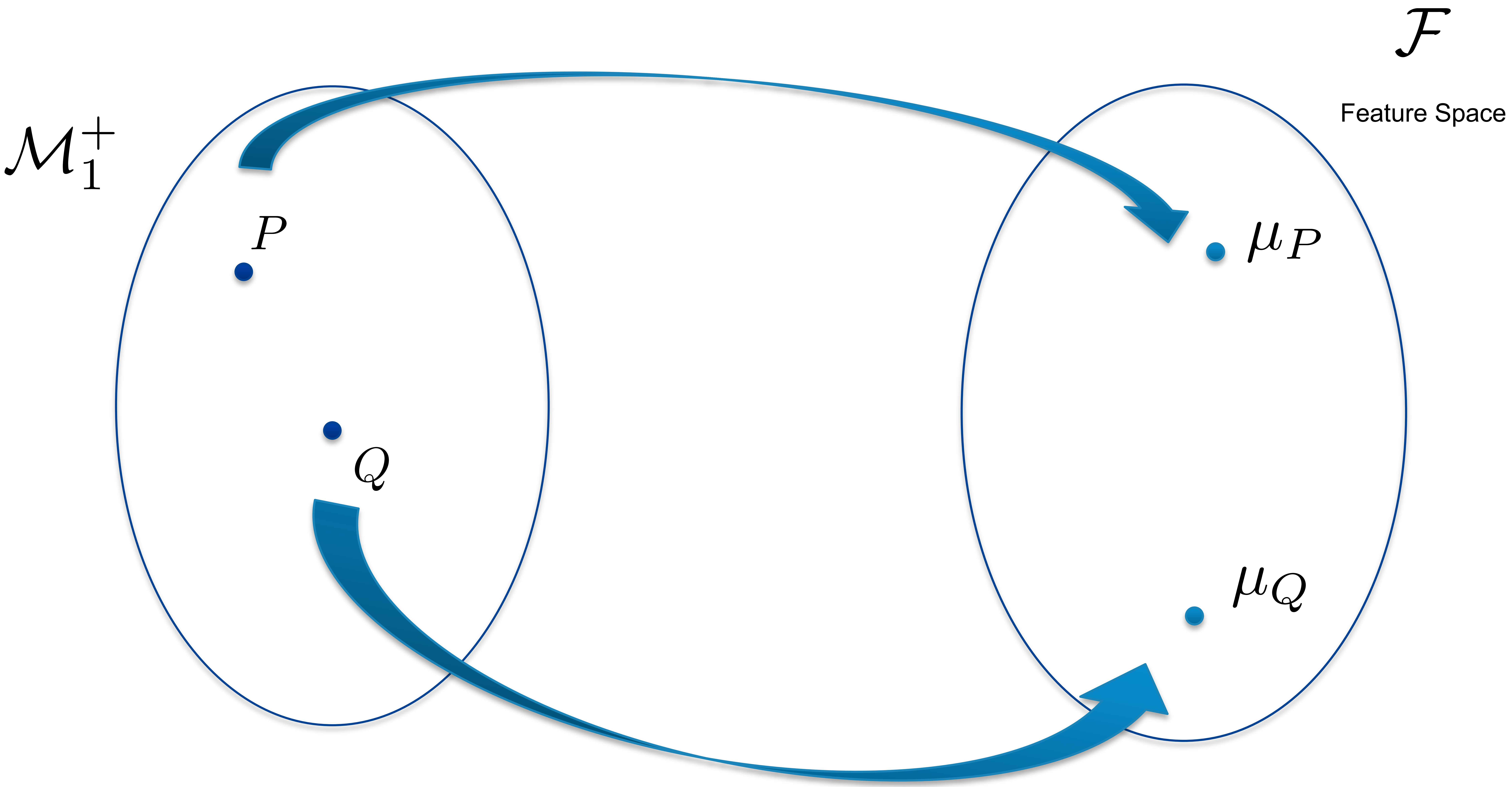
Kernel-embedding of probability distributions



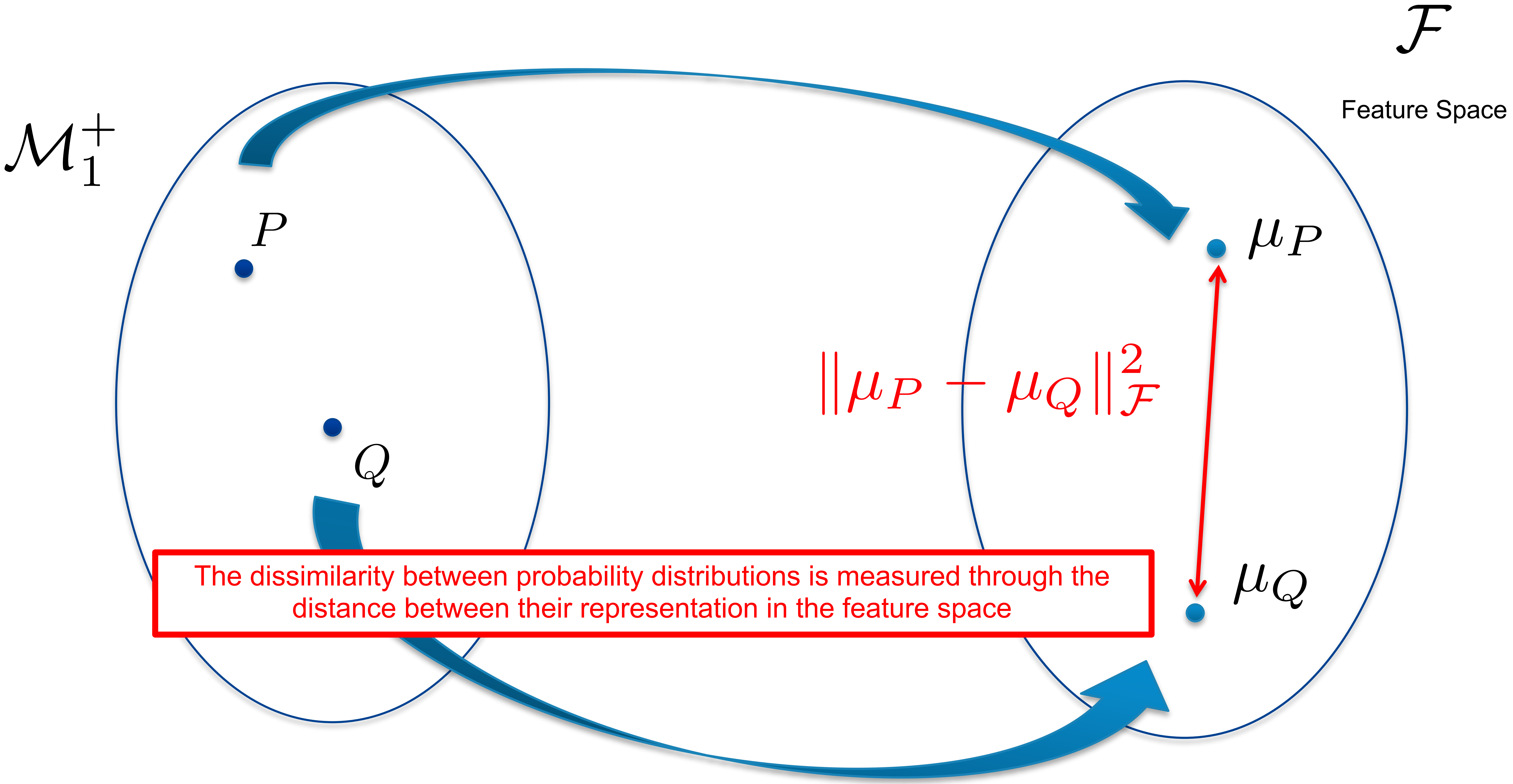
Option 1: work directly in the space of probability measures
Examples: KS, TV, KL, Hellinger, …

Option 2: represent probability measures with some features

Kernel-embedding of probability distributions

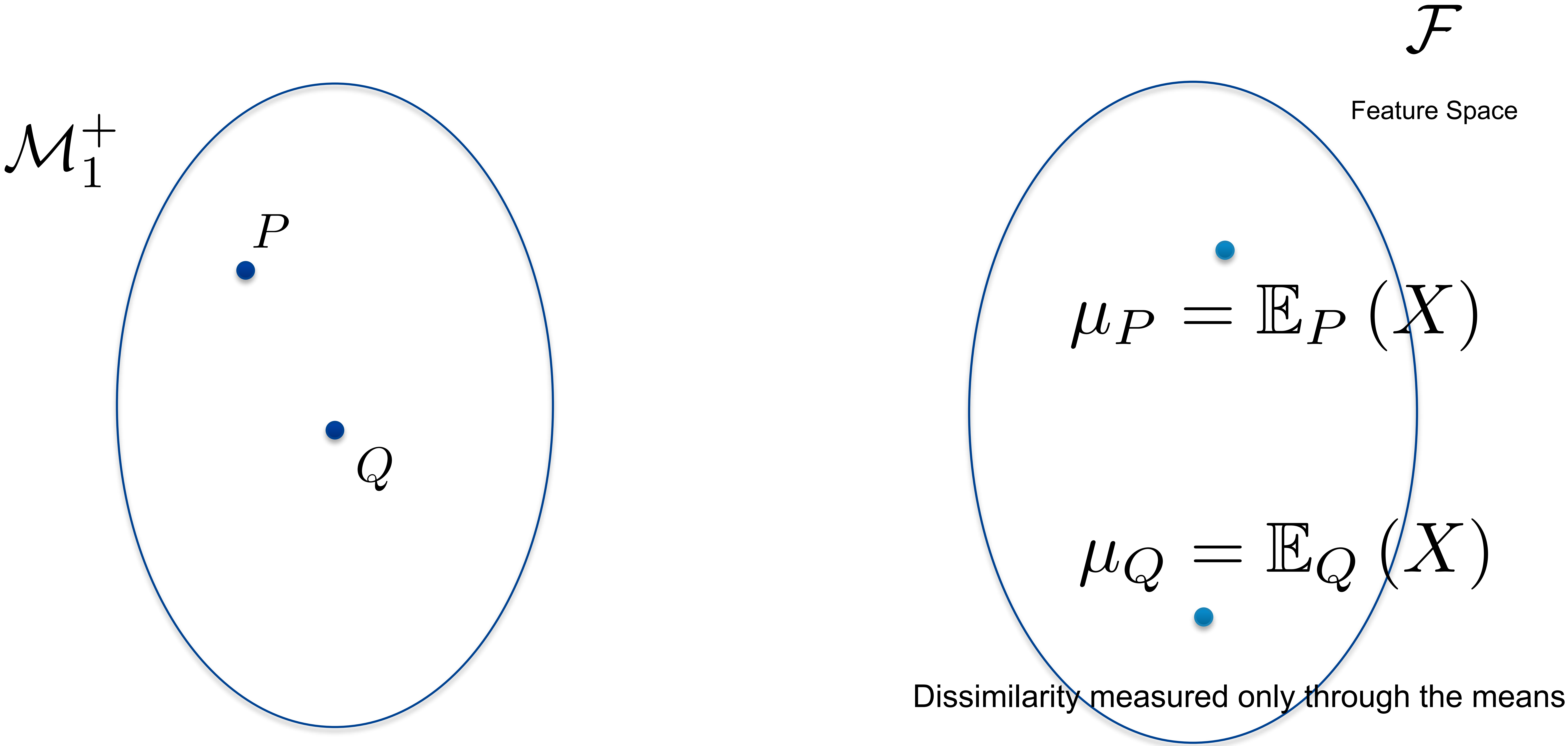


Kernel-embedding of probability distributions

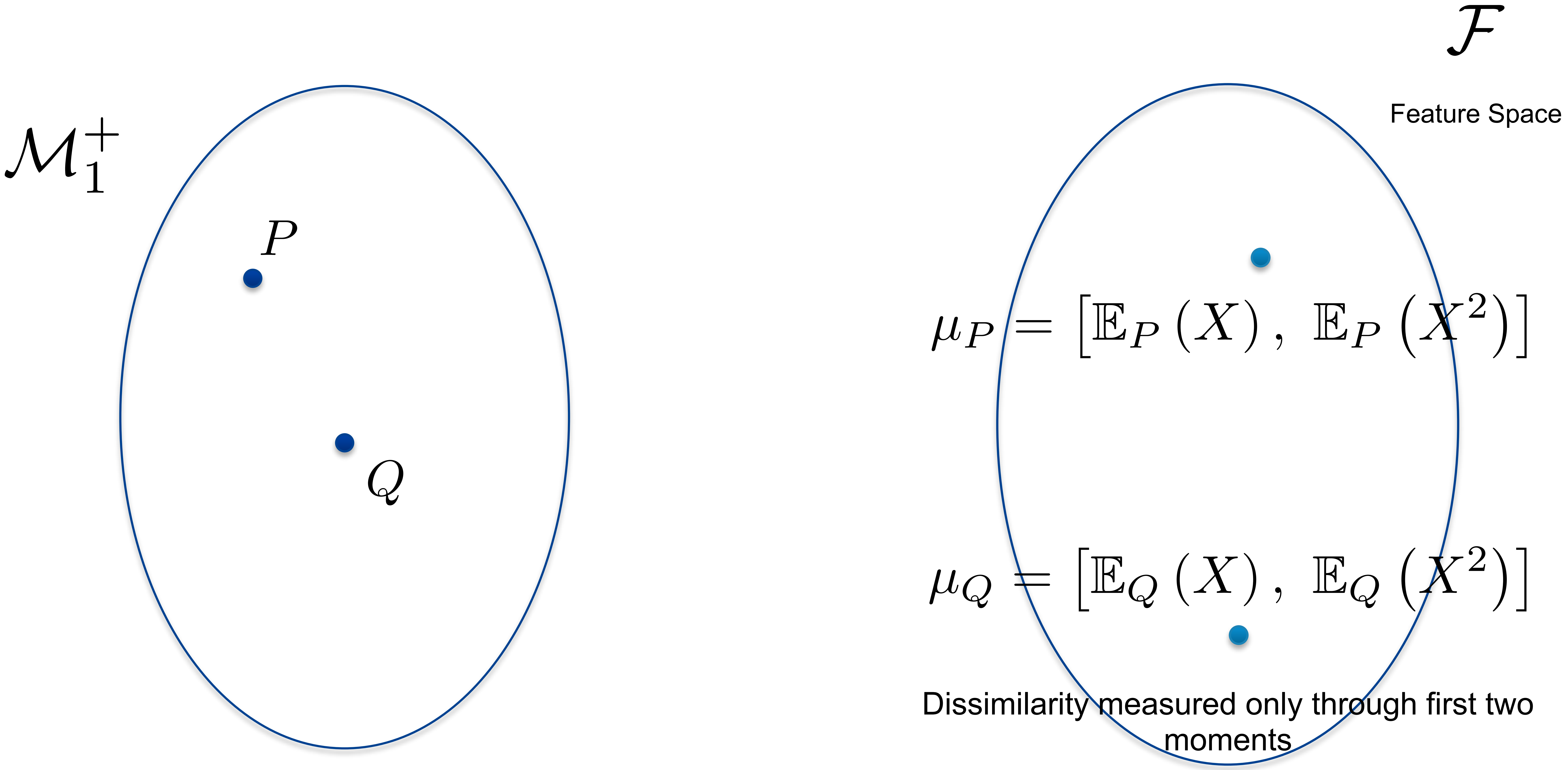


The dissimilarity between probability distributions is measured through the distance between their representation in the feature space

Kernel-embedding of probability distributions



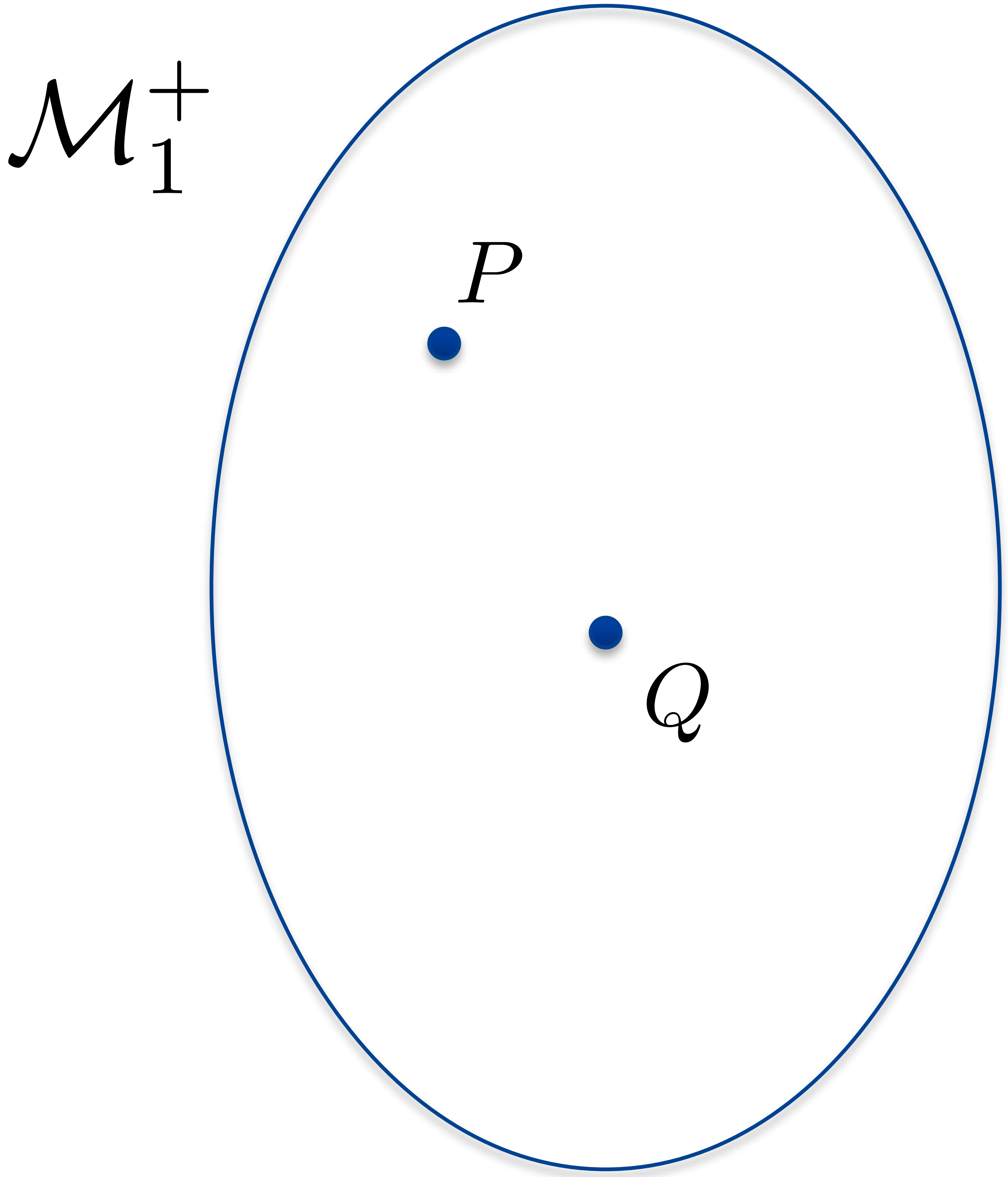
Kernel-embedding of probability distributions

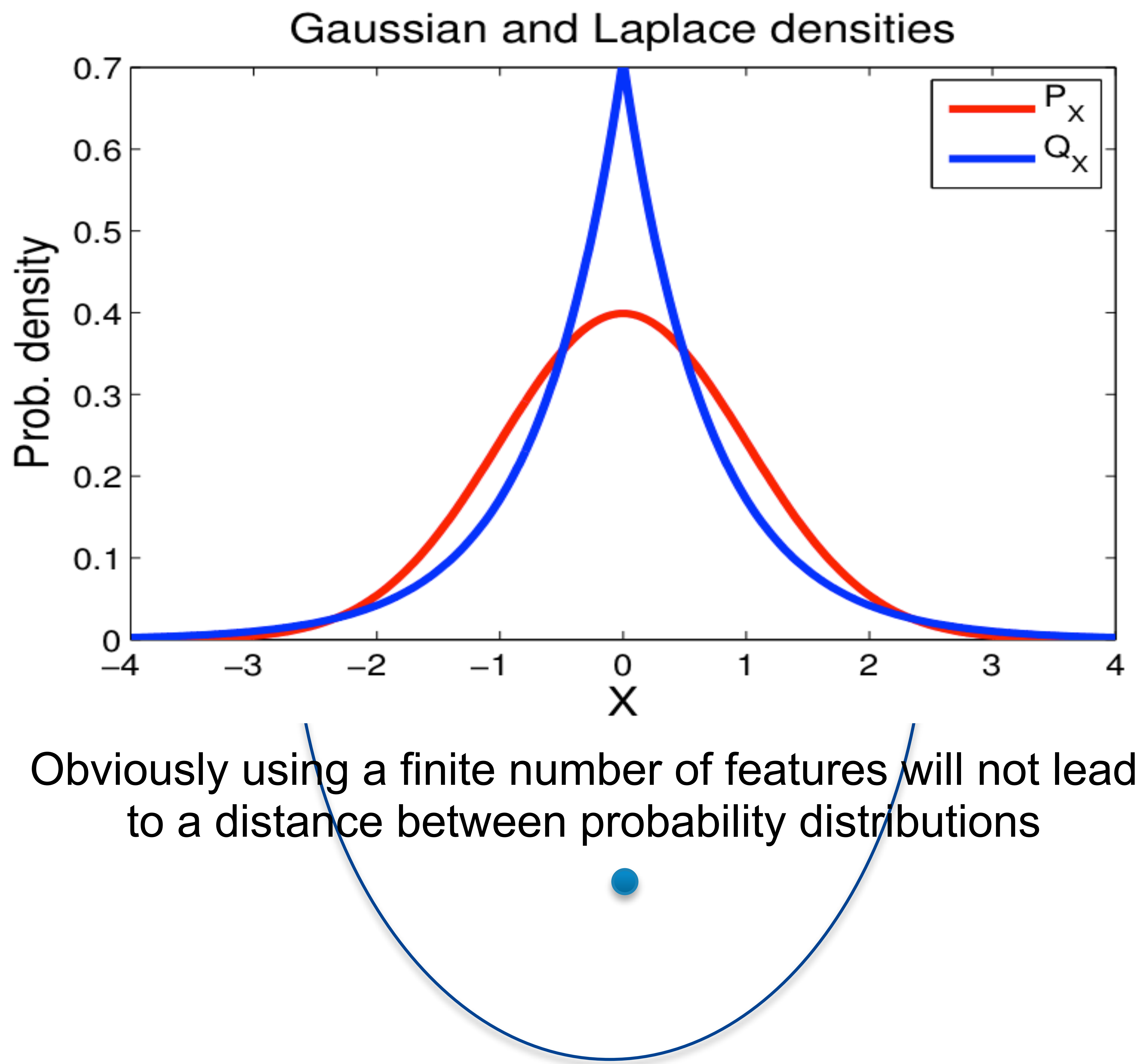


$$\mu_P = \left[\mathbb{E}_P(X),\ \mathbb{E}_P\left(X^2\right)\right]$$

$$\mu_Q = \left[\mathbb{E}_Q(X),\ \mathbb{E}_Q\left(X^2\right)\right]$$

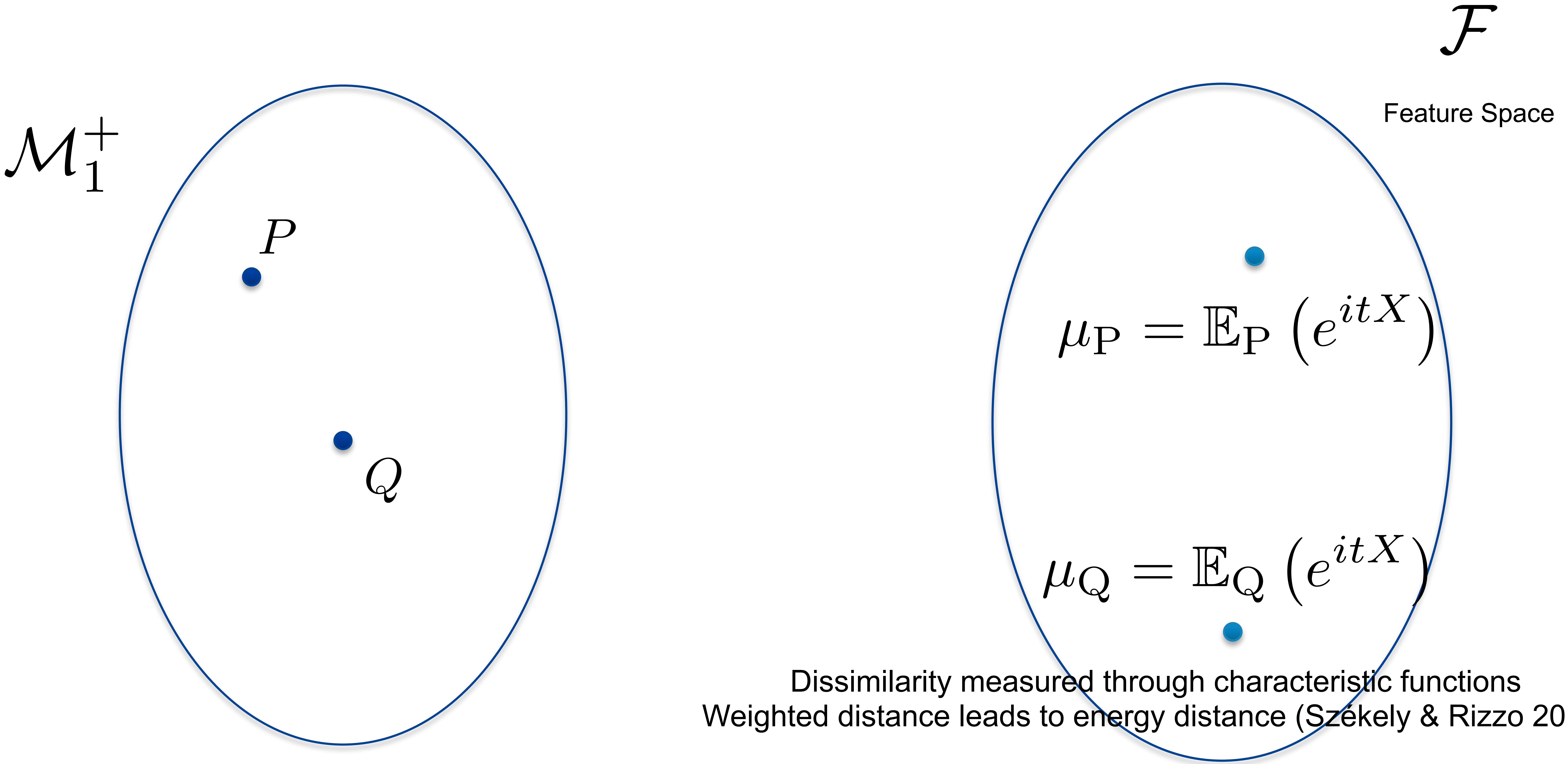
Kernel-embedding of probability distributions





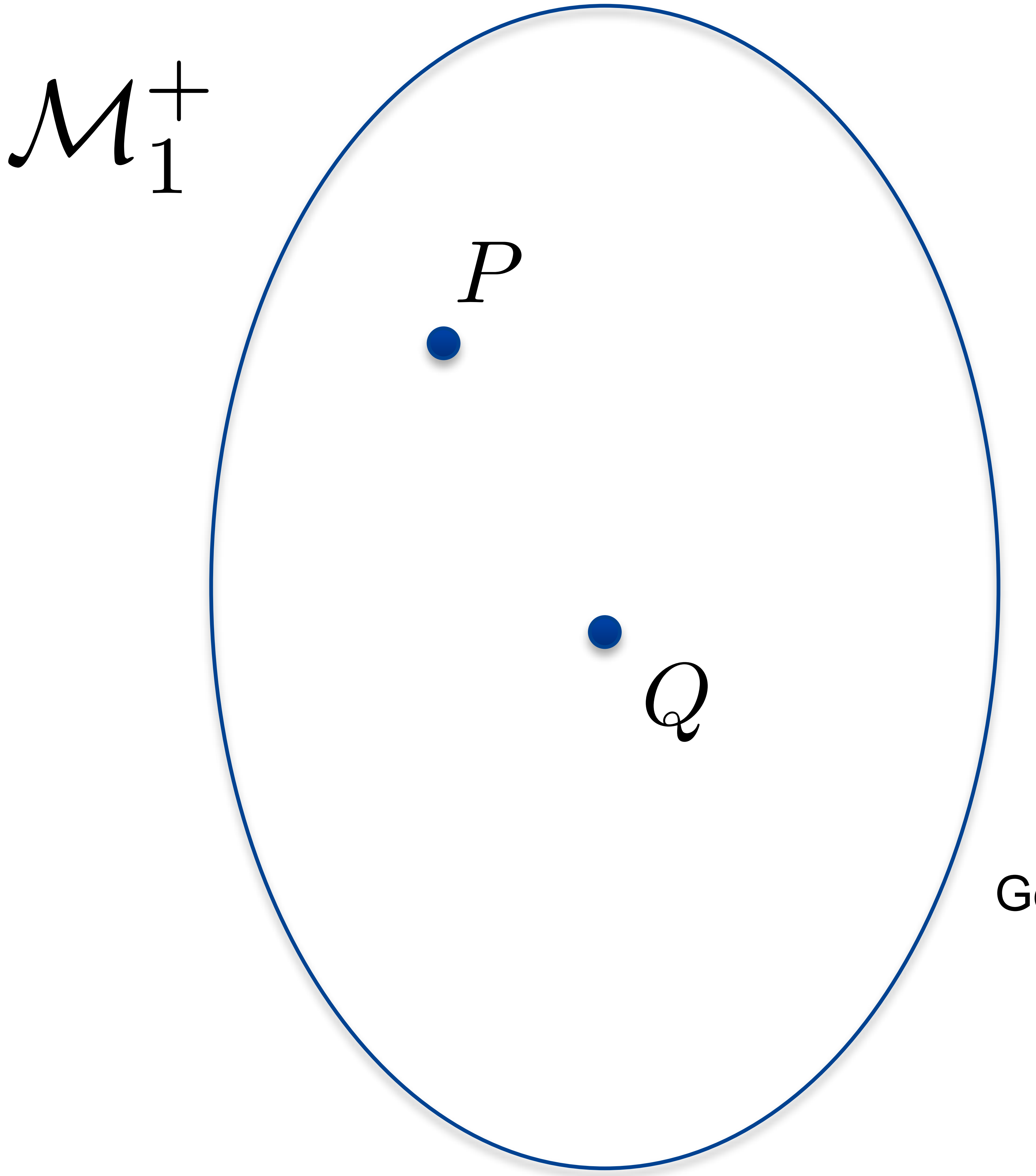
Obviously using a finite number of features will not lead to a distance between probability distributions

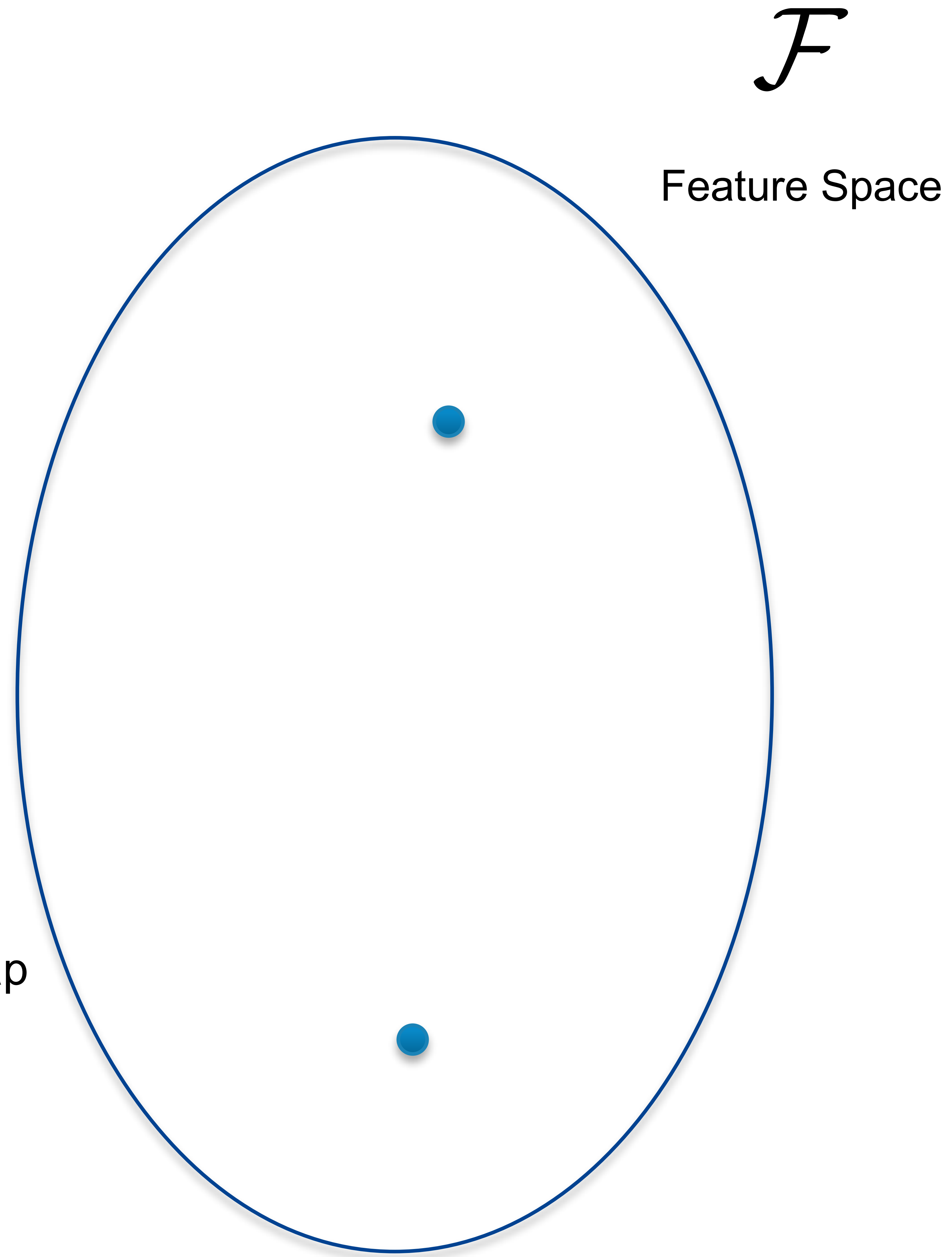
Kernel-embedding of probability distributions



Dissimilarity measured through characteristic functions
Weighted distance leads to energy distance (Székely & Rizzo 2013)

Kernel-embedding of probability distributions

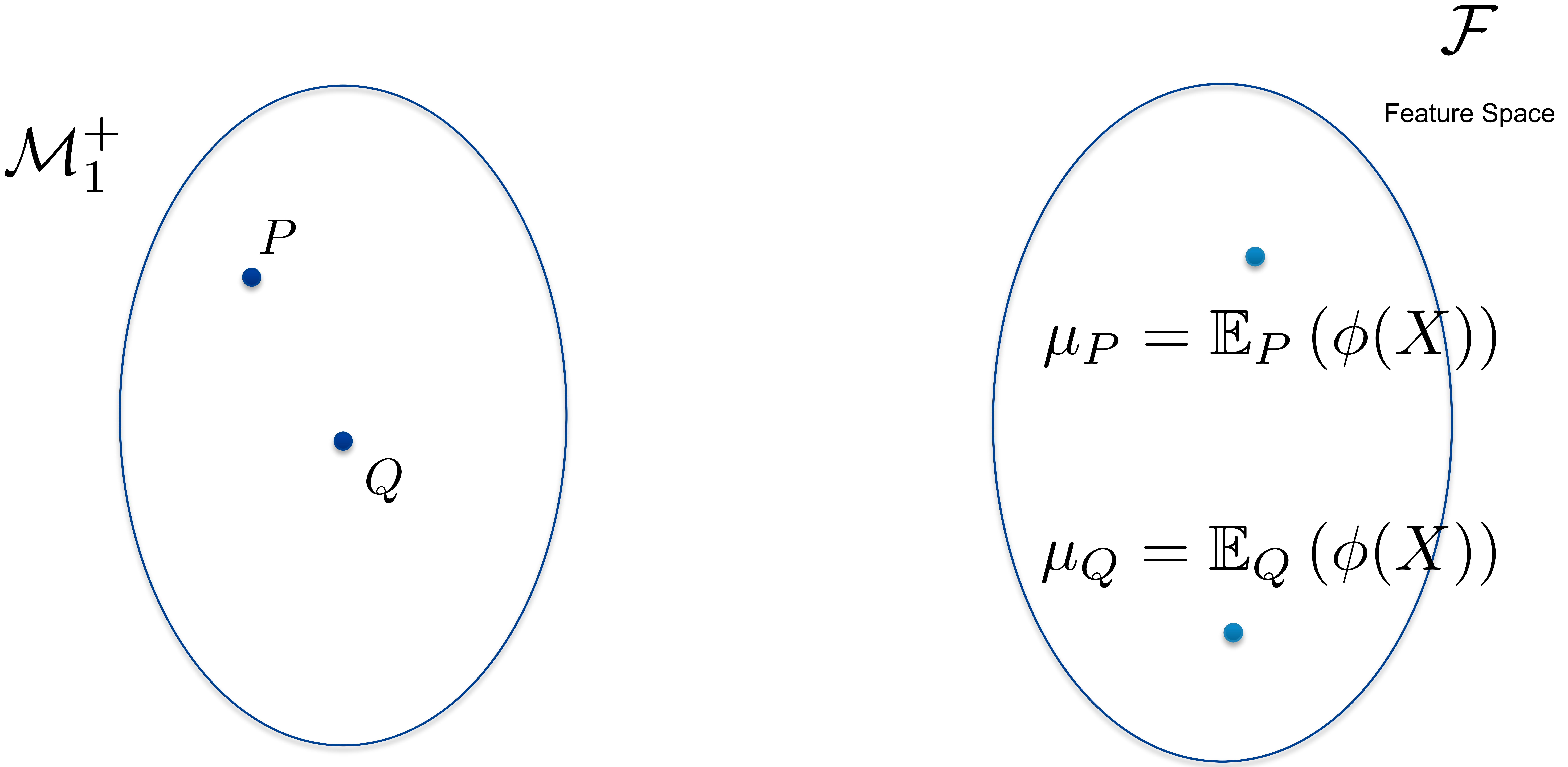




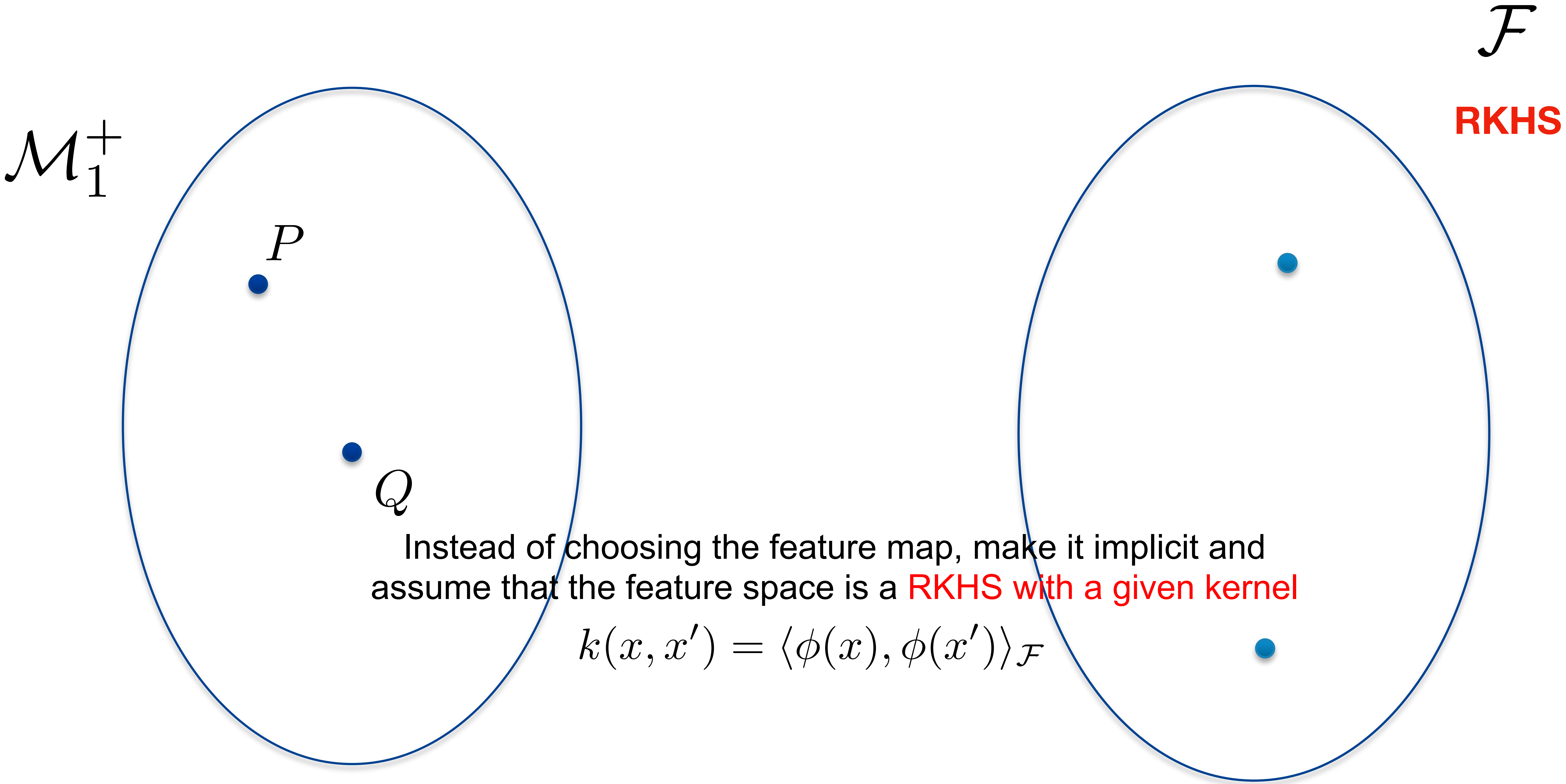
General setting: take a feature map

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

Kernel-embedding of probability distributions



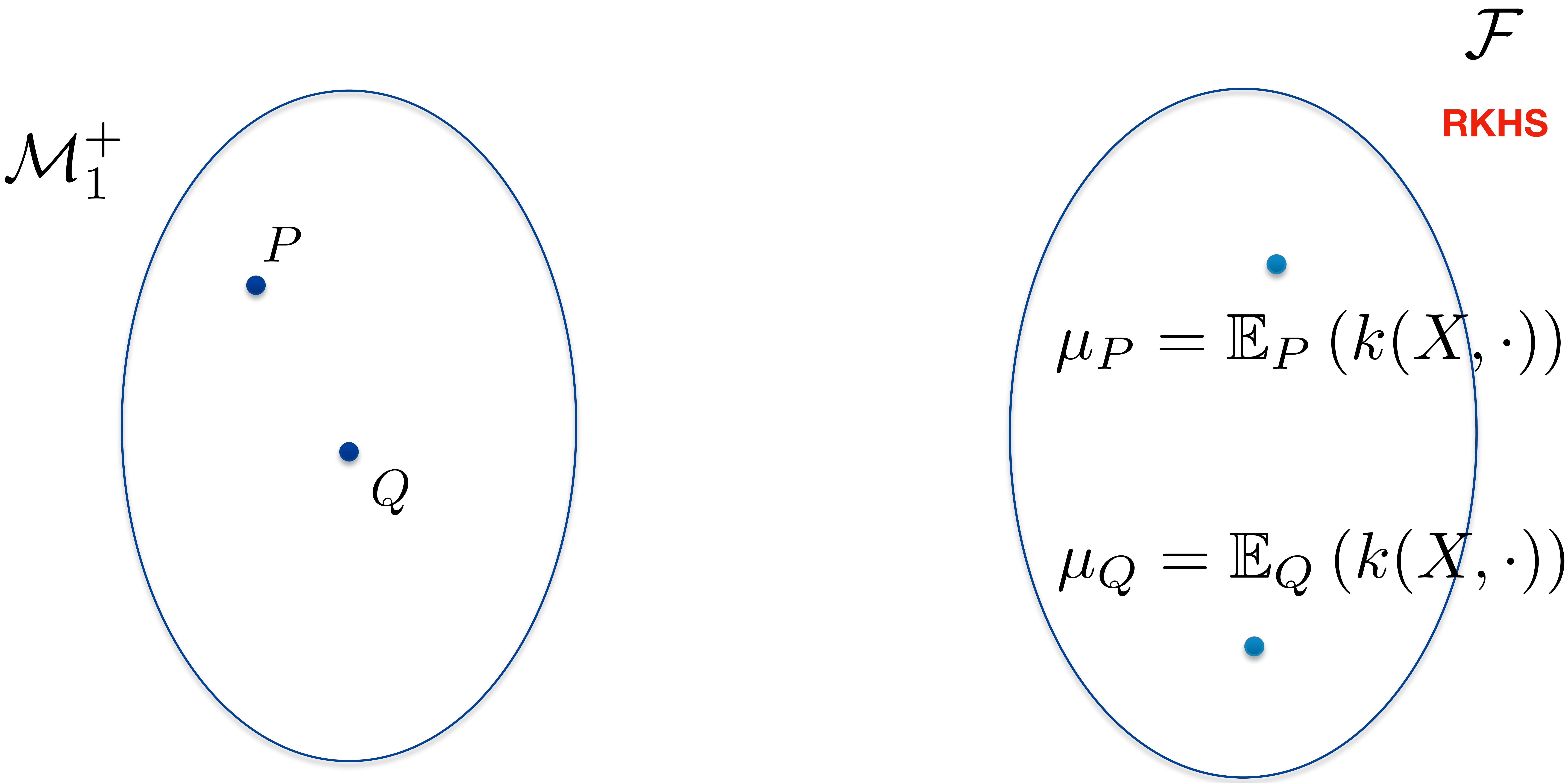
Kernel-embedding of probability distributions



Instead of choosing the feature map, make it implicit and assume that the feature space is a RKHS with a given kernel

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$$

Kernel-embedding of probability distributions



Kernel-embedding of probability distributions

The kernel mean embedding of a probability measure is defined as

$$\mu_{\mathrm{P}} = \mathbb{E}_{\xi \sim \mathrm{P}} k_{\mathcal{X}}(\xi, \cdot) = \int_{\mathcal{X}} k_{\mathcal{X}}(\xi, \cdot) d\mathrm{P}(\xi)$$

A distance between probability measures is then given by the **Maximum Mean Discrepancy**

$$\mathrm{MMD}(\mathrm{P}_1, \mathrm{P}_2) = \|\mu_{\mathrm{P}_1} - \mu_{\mathrm{P}_2}\|_{\mathcal{H}}$$

The reproducing property in the RKHS gives the central result

$$\mathrm{MMD}^2(\mathrm{P}_1, \mathrm{P}_2) = \mathbb{E}_{\xi,\xi'} k_{\mathcal{X}}(\xi, \xi') - 2\mathbb{E}_{\xi,\zeta} k_{\mathcal{X}}(\xi, \zeta) + \mathbb{E}_{\zeta,\zeta'} k_{\mathcal{X}}(\zeta, \zeta')$$

Smola et al. 2007, Song 2008, Song et al. 2009

Kernel-embedding of probability distributions

Advantages of this distance vs others

- Thanks to the RKHS, only involves **expectations of kernels**
- Less prone to the curse of dimensionality
- **Can easily handle structured objects** (curves, images, graphs, probability measures, sets) by using specific kernels

Kernel-embedding of probability distributions

Advantages of this distance vs others

- Thanks to the RKHS, only involves **expectations of kernels**
- Less prone to the curse of dimensionality
- **Can easily handle structured objects** (curves, images, graphs, probability measures, sets) by using specific kernels

See J. Pelamatti's talk

See N. Fellmann's talk

Kernel-embedding of probability distributions

Advantages of this distance vs others

- Thanks to the RKHS, only involves **expectations of kernels**
- Less prone to the curse of dimensionality
- **Can easily handle structured objects** (curves, images, graphs, probability measures, sets) by using specific kernels

For GSA, we will just plug-in this distance inside the general formula!

$$\mathcal{S}_l = \mathbb{E}_{X_l}\left(d(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right)$$

This means that we will define a kernel on the outputs
As a side effect, this gives a straightforward way to account for many output types in a computer code

Kernel-embedding of probability distributions

First-order

$$\begin{aligned}\mathcal{S}_l^{\mathrm{MMD}} &= \mathbb{E}_{X_l}\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|X_l}) \\ &= \mathbb{E}_{X_l}\mathbb{E}_{\xi,\xi'\sim\mathrm{P}_Y} k_{\mathcal{Y}}(\xi,\xi') - 2\mathbb{E}_{X_l}\mathbb{E}_{\xi\sim\mathrm{P}_Y,\zeta\sim\mathrm{P}_{Y|X_l}} k_{\mathcal{Y}}(\xi,\zeta) + \mathbb{E}_{X_l}\mathbb{E}_{\zeta,\zeta'\sim\mathrm{P}_{Y|X_l}} k_{\mathcal{Y}}(\zeta,\zeta') \\ &= \mathbb{E}_{X_l}\mathbb{E}_{\zeta,\zeta'\sim\mathrm{P}_{Y|X_l}} k_{\mathcal{Y}}(\zeta,\zeta') - \mathbb{E}_{\xi,\xi'\sim\mathrm{P}_Y} k_{\mathcal{Y}}(\xi,\xi')\end{aligned}$$

Group

$$\mathcal{S}_A^{\mathrm{MMD}} = \mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right) = \mathbb{E}_{\mathbf{X}_A}\mathbb{E}_{\zeta,\zeta'\sim\mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\zeta,\zeta') - \mathbb{E}_{\xi,\xi'\sim\mathrm{P}_Y} k_{\mathcal{Y}}(\xi,\xi')$$

D. 2016 & 2021, Barr & Rabitz 2022

Kernel-embedding of probability distributions

Example: stochastic simulator with 5 input variables

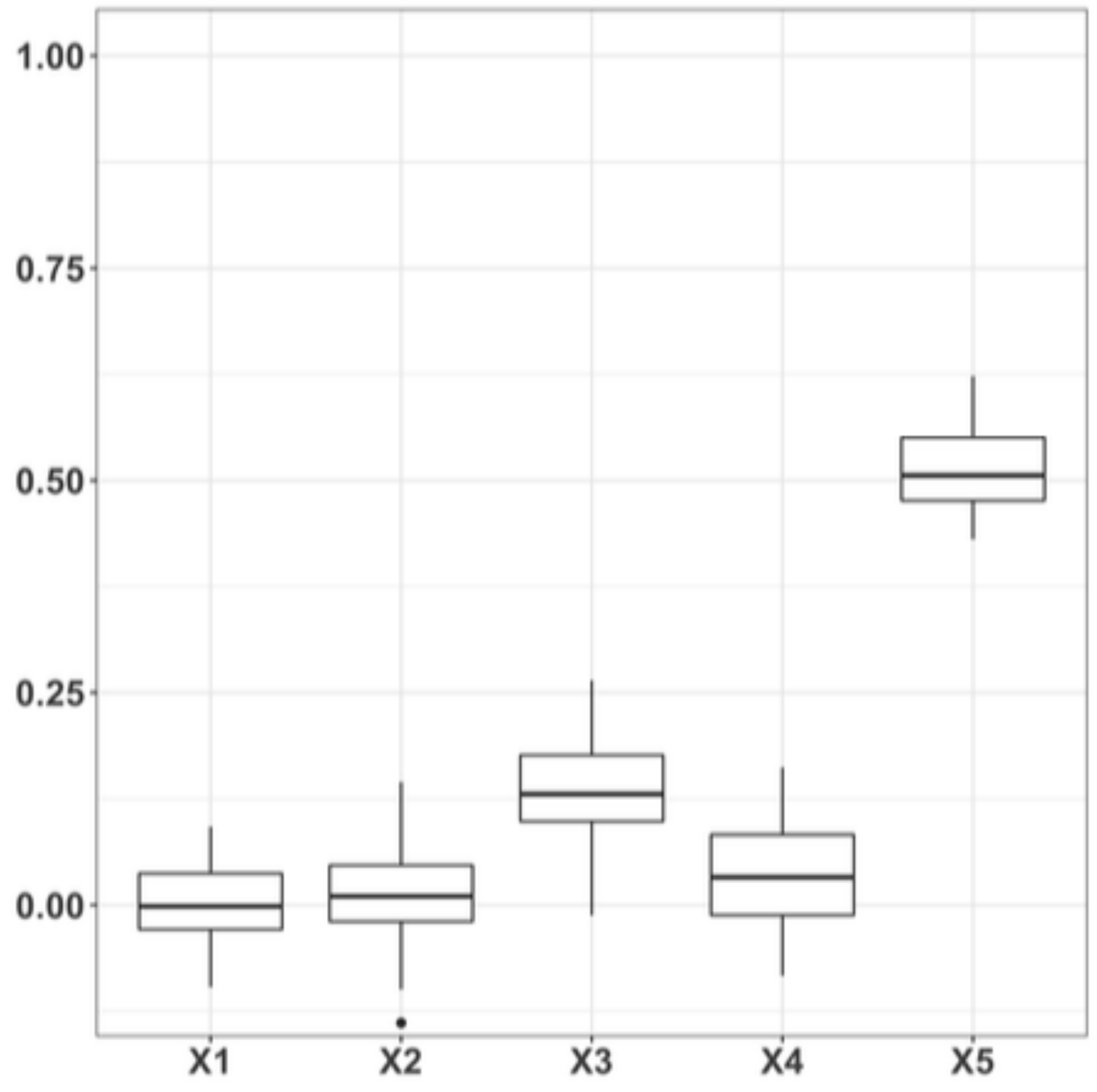
$$Y = (X_1 + 2X_2 + U_1)\sin(3X_3 - 4X_4 + N) + U_2 + 5X_5B + \sum_{i=1}^{5} iX_i$$

Input variables

« Internal » random variables responsible for code stochasticity

$X_1, \ldots, X_5 \sim \mathcal{U}(0,1)$ $\quad U_1 \sim \mathcal{U}(0,1),\ U_2 \sim \mathcal{U}(1,2),\ N \sim \mathcal{N}(0,1)\ \ B \sim \text{Bernoulli}(1/2)$

- **Use of a specific kernel to compare probability distributions (see J. Pelamatti's talk)**



(a) MMD first-order index

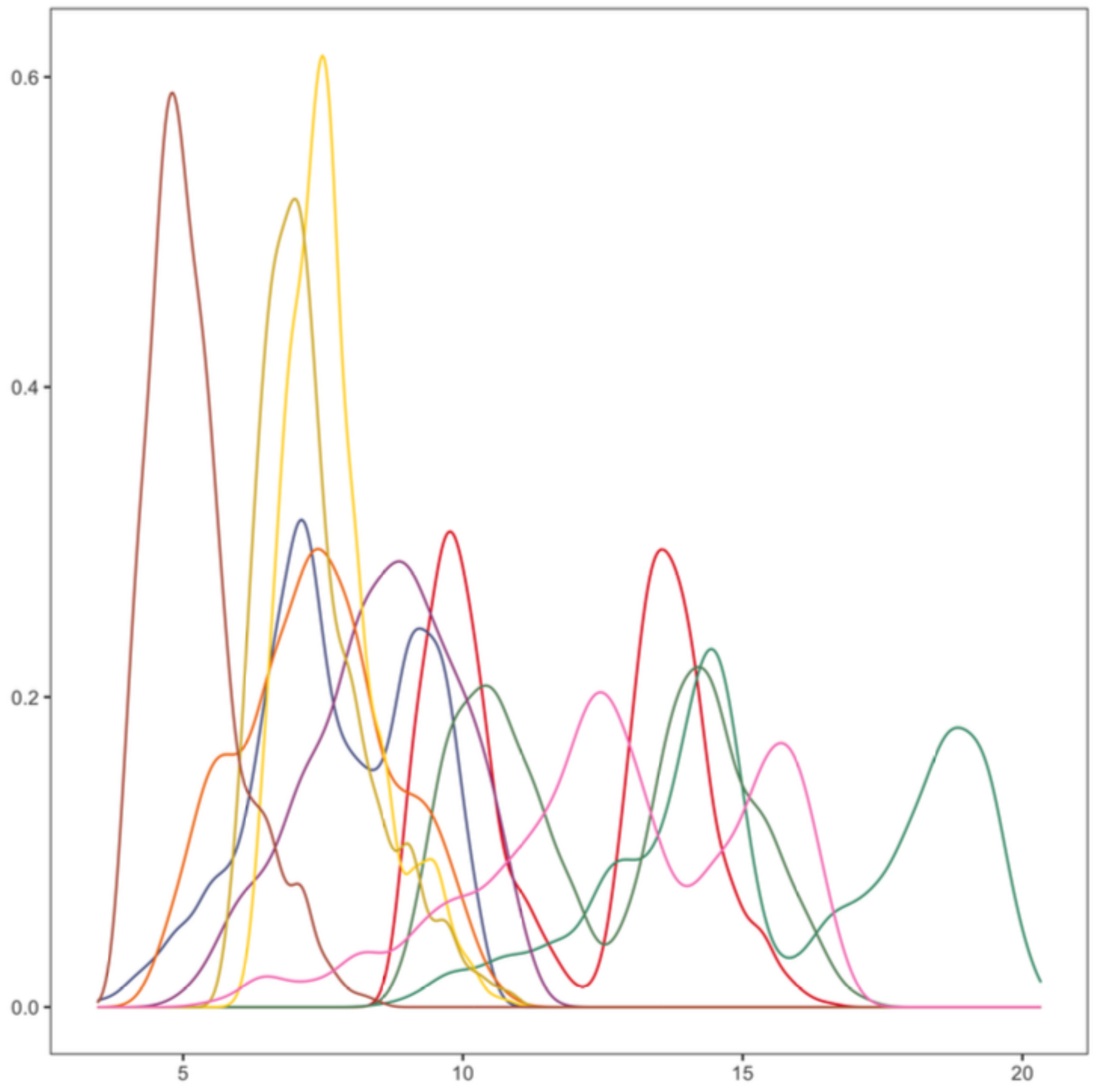


Figure 3: Stochastic simulator test case. Output probability distribution for 20 values of the input variables chosen at random. The distribution is estimated with a kernel-density estimator.

Kernel-embedding of probability distributions

Links with Sobol': if we use the vanilla dot product kernel $k_{\mathcal{Y}}(y, y') = yy'$

$$\begin{aligned} \mathcal{S}_A^{\mathrm{MMD}} &= \mathbb{E}_{\mathbf{X}_A} \left(\mathbb{E}_{\xi \sim \mathrm{P}_Y}(\xi) - \mathbb{E}_{\zeta \sim \mathrm{P}_{Y|\mathbf{X}_A}}(\zeta) \right)^2 \\ &= \mathbb{E}_{\mathbf{X}_A} \left(\mathbb{E}Y - \mathbb{E}(Y|\mathbf{X}_A) \right)^2 \\ &= \operatorname{Var} \mathbb{E}(Y|\mathbf{X}_A) \end{aligned}$$

Unnormalized Sobol'

Kernel-embedding of probability distributions

Links with Sobol': if we use the vanilla dot product kernel $k_{\mathcal{Y}}(y, y') = yy'$

$$\begin{aligned}
\mathcal{S}_A^{\mathrm{MMD}} &= \mathbb{E}_{\mathbf{X}_A}\left(\mathbb{E}_{\xi\sim \mathrm{P}_Y}(\xi) - \mathbb{E}_{\zeta\sim \mathrm{P}_{Y|\mathbf{X}_A}}(\zeta)\right)^2 \\
&= \mathbb{E}_{\mathbf{X}_A}\left(\mathbb{E}Y - \mathbb{E}(Y|\mathbf{X}_A)\right)^2 \\
&= \mathrm{Var}\,\mathbb{E}(Y|\mathbf{X}_A)
\end{aligned}$$

Unnormalized Sobol'

Links with Sobol': if Mercer's theorem holds

$$k_{\mathcal{Y}}(y, y') = \sum_{r=1}^{\infty} \phi_r(y)\phi_r(y') \Rightarrow$$

$$\begin{aligned}
\mathcal{S}_A^{\mathrm{MMD}} &= \sum_{r=1}^{\infty}\left\{\mathbb{E}_{\mathbf{X}_A}\mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_A}}\left(\phi_r(\xi)\phi_r(\xi')\right) - \mathbb{E}_{\zeta,\zeta'\sim \mathrm{P}}\left(\phi_r(\zeta)\phi_r(\zeta')\right)\right\} \\
&= \sum_{r=1}^{\infty}\left\{\mathbb{E}_{\mathbf{X}_A}\mathbb{E}\left(\phi_r(Y)|\mathbf{X}_A\right)^2 - \mathbb{E}\left(\phi_r(Y)\right)^2\right\} \\
&= \sum_{r=1}^{\infty}\mathrm{Var}\,\mathbb{E}\left(\phi_r(Y)|\mathbf{X}_A\right).
\end{aligned}$$

> **Aggregation of Sobol' indices on a (possibly) infinite number of nonlinear transformations of the output**

Kernel-embedding of probability distributions

Advantages of this distance vs others

- Thanks to the RKHS, only involves expectations of kernels
- Less prone to the curse of dimensionality
- Can easily handle structured objects (curves, images, graphs, probability measures, sets) by using specific kernels

- **Working in a RKHS gives access to orthogonal projections and decompositions**

Kernel-embedding of probability distributions

More importantly, we have an ANOVA-like decomposition !

Theorem 3 (ANOVA decomposition for MMD). *Under the same assumptions of Theorem 1 (in particular, the random vector* $\mathbf{X}$ *has independent components) and with Assumption 1, denote* $\mathrm{MMD}^2_{\mathrm{tot}} = \mathbb{E}k_{\mathcal{Y}}(Y,Y) - \mathbb{E}k_{\mathcal{Y}}(Y,Y')$ *where* Y' *is an independent copy of* Y*. Then the total MMD can be decomposed as*

$$\mathrm{MMD}^2_{\mathrm{tot}} = \sum_{A \subseteq \mathcal{P}_d} \mathrm{MMD}^2_A$$

where each term is given by

$$\mathrm{MMD}^2_A = \sum_{B \subset A} (-1)^{|A|-|B|} \mathbb{E}_{\mathbf{X}_B} \left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_B}) \right).$$

> **So we can define properly normalized MMD-based sensitivity indices**
> Proof is straightforward with Mercer’s theorem

Kernel-embedding of probability distributions

Definition 2 (MMD-based sensitivity indices). *In the frame of Theorem 3, let $A \subseteq \mathcal{P}_d$. The normalized MMD-based sensitivity index associated to a subset A of input variables is defined as*

$$S_A^{\mathrm{MMD}} = \frac{\mathrm{MMD}_A^2}{\mathrm{MMD}_{\mathrm{tot}}^2},$$

Impact of a subset alone

while the total MMD-based index associated to A is

$$S_A^{T,\mathrm{MMD}} = \sum_{B \subseteq \mathcal{P}_d,\, B \cap A \neq \emptyset} S_B^{\mathrm{MMD}} = 1 - \frac{\mathbb{E}_{\mathbf{X}_{-A}}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_{-A}})\right)}{\mathrm{MMD}_{\mathrm{tot}}^2}.$$

Impact of a subset through all its potential interactions with others

From Theorem 3, we have the fundamental identity providing the interpretation of MMD-based indices as percentage of the explained generalized variance $\mathrm{MMD}_{\mathrm{tot}}^2$:

$$\sum_{A \subseteq \mathcal{P}_d} S_A^{\mathrm{MMD}} = 1.$$

Interpretation as percentage

Kernel-embedding of probability distributions

New MMD-based sensitivity index

- **First moment-independent index with a decomposition**
- Can handle easily structured outputs
- Close generalization of Sobol' index, which is obtained as a particular case

Estimation

- We can easily recycle estimators proposed for Sobol' indices
- Monte-Carlo, Pick-freeze, Rank, k-NN
- See D. 2021 for details

Sensitivity analysis: our journey today

	Independent inputs		
	Sobol		Moment-independent
	1st order	Total order	Density-based
Beyond variance	✗	✗	✓
ANOVA (ranking)	✓	✓	✗
Screening	✗	✓	✗
Estimation (given data + small data)	✓	✗	✗
Can handle dependent inputs	✗	✗	✗
Can handle any output type	✗	✗	✗

Sensitivity analysis: our journey today

	Independent inputs				
	Sobol		Moment-independent		
	1st order	Total order	Density-based	1st order MMD	Total order MMD
Beyond variance	✗	✗	✓	✓	✓
ANOVA (ranking)	✓	✓	✗	✓	✓
Screening	✗	✓	✗	✗	✓
Estimation (given data + small data)	✓	✗	✗	✓	✗
Can handle dependent inputs	✗	✗	✗	✗	✗
Can handle any output type	✗	✗	✗	✓	✓

Kernel-based sensitivity analysis with ANOVA decomposition!

But cannot be used for screening (yet) and estimation as difficult as for Sobol’

Kernel-embedding of probability distributions

Remember our general GSA setting ?

$$\mathcal{S}_l = \mathbb{E}_{X_l}\left(d(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right)$$

Other point of view

$$\begin{aligned}
\mathcal{S}_l^{KL} &= \int p_{Y|X_l=x}(y)\ln\left(\frac{p_{Y|X_l=x}(y)}{p_Y(y)}\right)p_{X_l}(x)dxdy \\
&= \int \ln\left(\frac{p_{Y,X_l}(y,x)}{p_Y(y)p_{X_l}(x)}\right)p_{Y,X_l}(y,x)dxdy \\
&= \mathrm{MI}(X_l, Y)
\end{aligned}$$

> The KL-based index actually corresponds to the mutual information between one of the inputs and the output, i.e. a measure of their dependence

Kernel-embedding of probability distributions

The MMD strikes back

Other major use: testing independence of random vectors

$$\mathrm{MMD}^2(\mathrm{P}_{\mathbf{UV}}, \mathrm{P}_{\mathbf{U}} \otimes \mathrm{P}_{\mathbf{V}}) = \|\mu_{\mathrm{P}_{\mathbf{UV}}} - \mu_{\mathrm{P}_{\mathbf{U}}} \otimes \mu_{\mathrm{P}_{\mathbf{V}}}\|^2_{\mathcal{H}}$$

$$\begin{aligned} \mathrm{HSIC}(\mathbf{U}, \mathbf{V}) &= \mathrm{MMD}^2(\mathrm{P}_{\mathbf{UV}}, \mathrm{P}_{\mathbf{U}} \otimes \mathrm{P}_{\mathbf{V}}) \\ &= \mathbb{E}_{\mathbf{U},\mathbf{U}',\mathbf{V},\mathbf{V}'} k_{\mathcal{X}}(\mathbf{U}, \mathbf{U}') k_{\mathcal{Y}}(\mathbf{V}, \mathbf{V}') \\ &+ \mathbb{E}_{\mathbf{U},\mathbf{U}'} k_{\mathcal{X}}(\mathbf{U}, \mathbf{U}') \mathbb{E}_{\mathbf{V},\mathbf{V}'} k_{\mathcal{Y}}(\mathbf{V}, \mathbf{V}') \\ &- 2\mathbb{E}_{\mathbf{U},\mathbf{V}} \left[\mathbb{E}_{\mathbf{U}'} k_{\mathcal{X}}(\mathbf{U}, \mathbf{U}') \mathbb{E}_{\mathbf{V}'} k_{\mathcal{Y}}(\mathbf{V}, \mathbf{V}') \right] \end{aligned}$$

Gretton et al. 2005a,b

Many applications: goodness-of-fit, independence tests, feature selection, …

Kernel-embedding of probability distributions

HSIC-based sensitivity index

$$\mathcal{S}_A^{HS} = \mathrm{HSIC}(\mathbf{X}_A, Y)$$

- Already proposed with a hand-made normalization in D. 2015
- Detects independence, with small sample size → **Screening!**
- A kernel for the output just like for the MMD + **now a kernel for the inputs**

Screening can be achieved via statistical tests of independence (De Lozzo & Marrel 2016)

Sensitivity analysis: our journey today

	Independent inputs				
	Sobol		Moment-independent		
	1st order	Total order	Density-based	1st order MMD	Total order MMD
Beyond variance	✗	✗	✓	✓	✓
ANOVA (ranking)	✓	✓	✗	✓	✓
Screening	✗	✓	✗	✗	✓
Estimation (given data + small data)	✓	✗	✗	✓	✗
Can handle dependent inputs	✗	✗	✗	✗	✗
Can handle any output type	✗	✗	✗	✓	✓

Sensitivity analysis: our journey today

	Independent inputs					
	Sobol		Moment-independent			
	1st order	Total order	Density-based	1st order MMD	Total order MMD	HSIC
Beyond variance	✗	✗	✓	✓	✓	✓
ANOVA (ranking)	✓	✓	✗	✓	✓	✗
Screening	✗	✓	✗	✗	✓	✓
Estimation (given data + small data)	✓	✗	✗	✓	✗	✓
Can handle dependent inputs	✗	✗	✗	✗	✗	✓ *
Can handle any output type	✗	✗	✗	✓	✓	✓

Kernel-based sensitivity analysis that can be used for screening

But we have lost the ANOVA decomposition ☹️

* Note: they do not require independence to perform screening with statistical hypothesis tests

Kernel-embedding of probability distributions

But actually no, there is an ANOVA decomposition for HSIC

ANOVA-like decomposition for HSIC

Theorem 4 (ANOVA decomposition for HSIC). *Under the same assumptions of Theorem 1 (in particular, the random vector* $\mathbf{X}$ *has independent components) and with Assumptions 2 and 3, the HSIC dependence measure between* $\mathbf{X} = (X_1, \dots, X_d)$ *and* Y *can be decomposed as*

$$\mathrm{HSIC}(\mathbf{X}, Y) = \sum_{A \subseteq \mathcal{P}_d} \mathrm{HSIC}_A$$

where each term is given by

$$\mathrm{HSIC}_A = \sum_{B \subset A} (-1)^{|A|-|B|} \mathrm{HSIC}(\mathbf{X}_B, Y)$$

and $\mathrm{HSIC}(\mathbf{X}_B, Y)$ *is defined with a product RKHS* $\mathcal{H}_B = \mathcal{F}_B \times \mathcal{G}$ *with kernel* $k_B(\mathbf{x}_B, \mathbf{x}'_B) k_{\mathcal{Y}}(y, y') = \prod_{l \in B} (1 + k_l(x_l, x'_l)) \, k_{\mathcal{Y}}(y, y')$ *as in (10).*

Assumption on the kernels used for the inputs

> **So we can define properly normalized HSIC-based sensitivity indices**

> Proof relies on orthogonal decompositions in RKHS (see Appendix)

Kernel-embedding of probability distributions

But actually no, there is an ANOVA decomposition for HSIC

Definition 3 (HSIC-based sensitivity indices)**.** *In the frame of Theorem 4, let $A \subseteq \mathcal{P}_d$. The normalized HSIC-based sensitivity index associated to a subset A of input variables is defined as*

$$S_A^{\text{HSIC}} = \frac{\text{HSIC}_A}{\text{HSIC}\,(\mathbf{X}, Y)},$$

Impact of a subset alone

while the total HSIC-based index associated to A is

$$S_A^{T,\text{HSIC}} = \sum_{B \subseteq \mathcal{P}_d,\, B \cap A \neq \emptyset} S_B^{\text{HSIC}} = 1 - \frac{\text{HSIC}(\mathbf{X}_{-A}, Y)}{\text{HSIC}\,(\mathbf{X}, Y)}.$$

Impact of a subset through all its potential interactions with others

From Theorem 4, we have the fundamental identity providing the interpretation of HSIC-based indices as percentage of the explained HSIC dependence measure between $\mathbf{X} = (X_1, \ldots, X_d)$ and Y:

$$\sum_{A \subseteq \mathcal{P}_d} S_A^{\text{HSIC}} = 1.$$

Interpretation as percentage

Sensitivity analysis: our journey today

	Independent inputs					
	Sobol		Moment-independent			
	1st order	Total order	Density-based	1st order MMD	Total order MMD	HSIC
Beyond variance	✗	✗	✓	✓	✓	✓
ANOVA (ranking)	✓	✓	✗	✓	✓	✗
Screening	✗	✓	✗	✗	✓	✓
Estimation (given data + small data)	✓	✗	✗	✓	✗	✓
Can handle dependent inputs	✗	✗	✗	✗	✗	✓ *
Can handle any output type	✗	✗	✗	✓	✓	✓

* Note: they do not require independence to perform screening with statistical hypothesis tests

Sensitivity analysis: our journey today

	Independent inputs							
	Sobol		**Moment-independent**					
	1st order	Total order	Density-based	1st order MMD	Total order MMD	HSIC	1st order HSIC ANOVA	Total order HSIC ANOVA
Beyond variance	✗	✗	✓	✓	✓	✓	✓	✓
ANOVA (ranking)	✓	✓	✗	✓	✓	✗	✓	✓
Screening	✗	✓	✗	✗	✓	✓	✓	✓
Estimation (given data + small data)	✓	✗	✗	✓	✗	✓	✓	✓
Can handle dependent inputs	✗	✗	✗	✗	✗	✓ *	✗	✗
Can handle any output type	✗	✗	✗	✓	✓	✓	✓	✓

Kernel-based sensitivity analysis that can be used for screening and with an ANOVA decomposition

* Note: they do not require independence to perform screening with statistical hypothesis tests

Sensitivity analysis: our journey today

	Independent inputs							
	Sobol		Moment-independent					
	1st order	Total order	Density-based	1st order MMD	Total order MMD	HSIC	1st order HSIC ANOVA	Total order HSIC ANOVA
Beyond variance	✗	✗	✓	✓	✓	✓	✓	✓
ANOVA (ranking)	✓	✓	✗	✓	✓	✗	✓	✓
Screening	✗	✓	✗	✗	✓	✓	✓	✓
Estimation (given data + small data)	✓	✗	✗	✓	✗	✓	✓	✓
Can handle dependent inputs	✗	✗	✗	✗	✗	✓*	✗	✗
Can handle any output type	✗	✗	✗	✓	✓	✓	✓	✓

The last step is to discuss how we can handle dependent inputs

* Note: they do not require independence to perform screening with statistical hypothesis tests

HANDLING DEPENDENT INPUTS

Sensitivity analysis: dependent inputs

- **When inputs are dependent, a large consensus in ML is to use Shapley effects**

➡ The building blocks are Sobol' indices (variances of conditional expectations)

➡ We have a quantitative ranking via a decomposition (i.e. they sum to 1)

➡ But we are no longer able to measure interactions, since they are mixed with the dependence

➡ (However Shapley effects suffer from limitations, and recent research aims at improving them, see e.g. Herin et al. 2022)

Sensitivity analysis: Shapley effects

Definition 4 (Shapley effects (Shapley, 1953)). *For any $l = 1\ldots, d$, the Shapley effect of input X_l is given by*

$$Sh_l = \frac{1}{\operatorname{Var} Y}\frac{1}{p}\sum_{A\subseteq\mathcal{P}_d,\, A\not\ni l}\binom{p-1}{|A|}^{-1}\left\{\operatorname{Var}\mathbb{E}\left(Y|\mathbf{X}_{A\cup\{l\}}\right)-\operatorname{Var}\mathbb{E}\left(Y|\mathbf{X}_A\right)\right\}. \quad (14)$$

Moreover, we have the following decomposition

$$\sum_{l=1}^{p} Sh_l = 1.$$

Sensitivity analysis: our journey today

	Dependent inputs	
	Shapley	Moment-independent
	Shapley	HSIC
Beyond variance	✗	✓
ANOVA (ranking)	✓	✗
Screening	✓	✓
Estimation (given data + small data)	✗	✓
Can handle dependent inputs	✓	✓ *
Can handle any output type	✗	✓

* Note: they do not require independence to perform screening with statistical hypothesis tests

Sensitivity analysis: our journey today

	Dependent inputs	
	Shapley	**Moment-independent**
	Shapley	HSIC
Beyond variance	✗	✓
ANOVA (ranking)	✓	✗
Screening	✓	✓
Estimation (given data + small data)	✗	✓
Can handle dependent inputs	✓	✓ *
Can handle any output type	✗	✓

We will now try to recycle our previous kernel-based indices to improve this picture!

* Note: they do not require independence to perform screening with statistical hypothesis tests

Sensitivity analysis: Shapley effects

Definition 4 (Shapley effects (Shapley, 1953)). *For any $l = 1 \ldots, d$, the Shapley effect of input X_l is given by*

$$Sh_l = \frac{1}{\operatorname{Var} Y} \frac{1}{p} \sum_{A \subseteq \mathcal{P}_d,\, A \not\ni l} \binom{p-1}{|A|}^{-1} \left\{ \operatorname{Var} \mathbb{E}\left(Y | \mathbf{X}_{A \cup \{l\}}\right) - \operatorname{Var} \mathbb{E}\left(Y | \mathbf{X}_A\right) \right\}. \tag{14}$$

This definition corresponds to the Shapley value (Shapley, 1953)

$$\phi_l = \frac{1}{p} \sum_{A \subseteq \mathcal{P}_d,\, A \not\ni l} \binom{p-1}{|A|}^{-1} \left\{ \operatorname{val}\left(A \cup \{l\}\right) - \operatorname{val}\left(A\right) \right\}$$

with value function $\operatorname{val} : \mathcal{P}_d \to \mathbb{R}_+$ *equal to* $\operatorname{val}(A) = \operatorname{Var} \mathbb{E}\left(Y | \mathbf{X}_A\right) / \operatorname{Var} Y$. *Moreover, we have the following decomposition*

$$\sum_{l=1}^{p} Sh_l = 1.$$

The definition is general, and we have flexibility for the value function!

The only requirement is that the value function satisfies $\operatorname{val} : \mathcal{P}_d \to \mathbb{R}_+$ such that $\operatorname{val}(\emptyset) = 0$.

Sensitivity analysis: Shapley effects

Definition 5 (Kernel-embedding Shapley effects). *For any $l = 1 \dots, d$, we define*

(a) The MMD-Shapley effect

$$Sh_l^{\mathrm{MMD}} = \frac{1}{\mathrm{MMD}^2_{\mathrm{tot}}} \frac{1}{p} \sum_{A \subseteq \mathcal{P}_d,\, A \not\ni l} \binom{p-1}{|A|}^{-1} \Big\{ \mathbb{E}_{\mathbf{X}_{A\cup\{l\}}} \left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_{A\cup\{l\}}}) \right) - \mathbb{E}_{\mathbf{X}_A} \left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A}) \right) \Big\}$$

provided Assumption 1 holds.

(b) The HSIC-Shapley effect

$$Sh_l^{\mathrm{HSIC}} = \frac{1}{\mathrm{HSIC}\,(\mathbf{X}, Y)} \frac{1}{p} \sum_{A \subseteq \mathcal{P}_d,\, A \not\ni l} \binom{p-1}{|A|}^{-1} \Big\{ \mathrm{HSIC}\left(\mathbf{X}_{A\cup\{l\}}, Y\right) - \mathrm{HSIC}\left(\mathbf{X}_A, Y\right) \Big\}$$

provided Assumptions 2 and 3 hold.

MMD

We plug the kernel-based indices

HSIC

Sensitivity analysis: our journey today

	Dependent inputs	
	Shapley	**Moment-independent**
	Shapley	HSIC
Beyond variance	✗	✓
ANOVA (ranking)	✓	✗
Screening	✓	✓
Estimation (given data + small data)	✗	✓
Can handle dependent inputs	✓	✓ *
Can handle any output type	✗	✓

* Note: they do not require independence to perform screening with statistical hypothesis tests

Sensitivity analysis: our journey today

	Dependent inputs			
	Shapley	**Moment-independent**		
	Shapley	HSIC	MMD-Shapley	HSIC-Shapley
Beyond variance	✗	✓	✓	✓
ANOVA (ranking)	✓	✗	✓	✓
Screening	✓	✓	✓	✓
Estimation (given data + small data)	✗	✓	✗	✓
Can handle dependent inputs	✓	✓ *	✓	✓
Can handle any output type	✗	✓	✓	✓

* Note: they do not require independence to perform screening with statistical hypothesis tests

Sensitivity analysis: our journey today

	Dependent inputs				
	Shapley	**Moment-independent**			
	Shapley	HSIC	MMD-Shapley	HSIC-Shapley	
Beyond variance	✗	✓	✓	✓
ANOVA (ranking)	✓	✗	✓	✓
Screening	✓	✓	✓	✓
Estimation (given data + small data)	✗	✓	✗	✓
Can handle dependent inputs	✓	✓*	✓	✓
Can handle any output type	✗	✓	✓	✓

MMD Shapley is to Shapley what MMD was to Sobol'

* Note: they do not require independence to perform screening with statistical hypothesis tests

Sensitivity analysis: our journey today

	Dependent inputs			
	Shapley	**Moment-independent**		
	Shapley	HSIC	MMD-Shapley	HSIC-Shapley
Beyond variance	✗	✓	✓	✓
ANOVA (ranking)	✓	✗	✓	✓
Screening	✓	✓	✓	✓
Estimation (given data + small data)	✗	✓	✗	✓
Can handle dependent inputs	✓	✓ *	✓	✓
Can handle any output type	✗	✓	✓	✓

MMD Shapley is to Shapley what MMD was to Sobol’

HSIC-Shapley seems to have the most potential

* Note: they do not require independence to perform screening with statistical hypothesis tests

Conclusions & Perspectives

Kernel-based sensitivity analysis seems to have the potential to answer several practical needs

- ANOVA decomposition just like Sobol'
- Screening at low cost, with given data
- Can handle a ton of (complicated) outputs
- **Most of them are now available in the sensitivity package!**

Conclusions & Perspectives

Kernel-based sensitivity analysis seems to have the potential to answer several practical needs

- ANOVA decomposition just like Sobol'
- Screening at low cost, with given data
- Can handle a ton of (complicated) outputs
- **Most of them are now available in the sensitivity package!**

But there is a catch

- The complexity is reported on the choice of the kernel(s)
  - ✓ There is a vast literature on this problem though
- Interpretation of these indices is less straightforward and natural when compared to Sobol'
  - This means we have still work to do from a theoretical and practical point of view (see e.g. G. Sarazin's postdoc results in ANR Samurai project)

Sensitivity analysis: our journey today

	Independent inputs								Dependent inputs			
	Sobol		Moment-independent						Shapley	Moment-independent		
	1st order	Total order	Density-based	1st order MMD	Total order MMD	HSIC	1st order HSIC ANOVA	Total order HSIC ANOVA	Shapley	HSIC	MMD-Shapley	HSIC-Shapley
Beyond variance	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓
ANOVA (ranking)	✓	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓	✓
Screening	✗	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓
Estimation (given data + small data)	✓	✗	✗	✓	✗	✓	✓	✓				✓
Can handle dependent inputs	✗	✗	✗	✗	✗	✓	✗	✗				✓
Can handle any output type	✗	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓

More work needed to better understand these indices

References

Barr, J., & Rabitz, H. (2022). A Generalized Kernel Method for Global Sensitivity Analysis. *SIAM/ASA Journal on Uncertainty Quantification*, *10*(1), 27-54.

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.

Chastaing, G., Gamboa, F., Prieur, C. et al. (2012), ‘Generalized hoeffding-sobol decomposition for dependent variables - application to sensitivity analysis’, Electronic Journal of Statistics 6, 2420–2448.

Cuturi, M. (2011), Fast global alignment kernels, in ‘Proceedings of the 28th international conference on machine learning (ICML-11)’, pp. 929–936.

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

Da Veiga, S. (2021), ‘Kernel-based anova decomposition and shapley effects - application to global sensitivity analysis’ , https://arxiv.org/abs/2101.05487.

Da Veiga, S., Gamboa, F., Iooss, B., & Prieur, C. (2021). *Basics and Trends in Sensitivity Analysis: Theory and Practice in R*. Society for Industrial and Applied Mathematics.

Fort, J.-C., Klein, T. and Rachdi, N. (2016), ‘New sensitivity analysis subordinated to a contrast’, Communications in Statistics-Theory and Methods 45(15), 4349–4364.

Gretton, A., Bousquet, O., Smola, A. and Scholkopf, B. (2005a), Measuring statistical dependence with hilbert-schmidt norms, in S. Jain, H. Simon and E. Tomita, eds, ‘Algorithmic Learning Theory’, Vol. 3734 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 63–77.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O. and Scholkopf, B. (2005b), ‘Kernel methods for measuring independence’, The Journal of Machine Learning Research 6, 2075–2129.

Hoeffding, W. (1948), ‘A class of statistics with asymptotically normal distributions’, Annals of Mathematical Statistics 19, 293–325.

Idrissi, M. I., Chabridon, V., and Iooss, B. (2021), ‘Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs’, *arXiv preprint arXiv:2101.08083*.

Iooss, B. and Prieur, C. (2019), ‘Shapley effects for sensitivity analysis with dependent inputs: comparisons with Sobol’ indices, numerical estimation and applications’, International Journal for Uncertainty Quantification 9, 493–514.

Kuo, F., Sloan, I., Wasilkowski, G. and Wozniakowski, H. (2010), ‘On decompositions of multivari- ate functions’, Mathematics of computation 79(270), 953–966.

Mara, T. A., Tarantola, S. and Annoni, P. (2015), ‘Non-parametric methods for global sensitivity analysis of model output with dependent inputs’, Environmental modelling & software 72, 173– 183.

Owen, A. (2014), ‘Sobol’ indices and Shapley value’, SIAM/ASA Journal on Uncertainty Quantification 2, 245–251.

Rahman, S. (2016), ‘The f-sensitivity index’, SIAM/ASA Journal on Uncertainty Quantification 4(1), 130–162.

Shapley, L. (1953), A value for n-persons game, in H. Kuhn and A. Tucker, eds, ‘Contributions to the theory of games II, Annals of mathematic studies’, Princeton University Press, Princeton, NJ.

Song, E., Nelson, B. L. and Staum, J. (2016), ‘Shapley effects for global sensitivity analysis: Theory and computation’, SIAM/ASA Journal on Uncertainty Quantification 4(1), 1060–1083.

Song, L. (2008), Learning via Hilbert Space Embedding of Distributions, PhD thesis, University of Sydney.

Székely, G. J., & Rizzo, M. L. (2013), ‘Energy statistics: A class of statistics based on distances’, *Journal of statistical planning and inference*, *143*(8), 1249-1272.

APPENDIX

Kernel-embedding of probability distributions for GSA: HSIC

HSIC-based sensitivity index

$$\mathcal{S}_A^{HS} = \mathrm{HSIC}(\mathbf{X}_A, Y)$$

- Already proposed with a hand-made normalization in D. 2015
- Works very well for screening, with small sample size

But it actually exhibits an ANOVA decomposition too

Assumption 3. *The reproducing kernel $k_{\mathcal{X}}$ of $\mathcal{F}$ is of the form*

$$k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} \left(1 + k_l(x_l, x_l')\right) \tag{10}$$

where for each $l = 1, \ldots, d$, $k_l(\cdot, \cdot)$ is the reproducing kernel of a RKHS $\mathcal{F}_l$ of real functions depending only on variable x_l and such that $1 \notin \mathcal{F}_l$.
In addition, for all $l = 1, \ldots, d$ and $\forall x_l \in \mathcal{X}_l$, we have

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0. \tag{11}$$

Kernel-embedding of probability distributions for GSA: HSIC

HSIC-based sensitivity index

$$\mathcal{S}_A^{HS} = \mathrm{HSIC}(\mathbf{X}_A, Y)$$

- Already proposed with a hand-made normalization in D. 2015
- Works very well for screening, with small sample size

But it actually exhibits an ANOVA decomposition too

Assumption 3. *The reproducing kernel $k_{\mathcal{X}}$ of $\mathcal{F}$ is of the form*

Product kernel

$$k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} \left(1 + k_l(x_l, x_l')\right) \tag{10}$$

where for each $l = 1, \ldots, d$, $k_l(\cdot,\cdot)$ is the reproducing kernel of a RKHS $\mathcal{F}_l$ of real functions depending only on variable x_l and such that $1 \notin \mathcal{F}_l$.
In addition, for all $l = 1, \ldots, d$ and $\forall x_l \in \mathcal{X}_l$, we have

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0. \tag{11}$$

Kernel-embedding of probability distributions for GSA: HSIC

HSIC-based sensitivity index

$$\mathcal{S}_A^{HS} = \mathrm{HSIC}(\mathbf{X}_A, Y)$$

- Already proposed with a hand-made normalization in D. 2015
- Works very well for screening, with small sample size

But it actually exhibits an ANOVA decomposition too

Assumption 3. *The reproducing kernel $k_{\mathcal{X}}$ of $\mathcal{F}$ is of the form*

Product kernel

$$k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} \left(1 + k_l(x_l, x_l')\right) \tag{10}$$

where for each $l = 1, \ldots, d$, $k_l(\cdot, \cdot)$ is the reproducing kernel of a RKHS $\mathcal{F}_l$ of real functions depending only on variable x_l and such that $1 \notin \mathcal{F}_l$. **Without constant functions**
In addition, for all $l = 1, \ldots, d$ and $\forall x_l \in \mathcal{X}_l$, we have

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0. \tag{11}$$

Kernel-embedding of probability distributions for GSA: HSIC

HSIC-based sensitivity index

$$\mathcal{S}_A^{HS} = \mathrm{HSIC}(\mathbf{X}_A, Y)$$

- Already proposed with a hand-made normalization in D. 2015
- Works very well for screening, with small sample size

But it actually exhibits an ANOVA decomposition too

Assumption 3. *The reproducing kernel $k_{\mathcal{X}}$ of $\mathcal{F}$ is of the form*

Product kernel

$$k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} \left(1 + k_l(x_l, x_l')\right) \tag{10}$$

where for each $l = 1, \ldots, d$, $k_l(\cdot, \cdot)$ is the reproducing kernel of a RKHS $\mathcal{F}_l$ of real functions depending only on variable x_l and such that $1 \notin \mathcal{F}_l$. **Without constant functions**
In addition, for all $l = 1, \ldots, d$ and $\forall x_l \in \mathcal{X}_l$, we have

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0. \tag{11}$$

Zero-mean kernel

Kernel-embedding of probability distributions for GSA: HSIC

HSIC-based sensitivity index

$$\mathcal{S}_A^{HS} = \mathrm{HSIC}(\mathbf{X}_A, Y)$$

- Already proposed with a hand-made normalization in D. 2015
- Works very well for screening, with small sample size

But it actually exhibits an ANOVA decomposition too

Needed to get orthogonality inside the RKHS

Assumption 3. *The reproducing kernel $k_{\mathcal{X}}$ of $\mathcal{F}$ is of the form*

Product kernel

$$k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} \left(1 + k_l(x_l, x_l')\right) \tag{10}$$

where for each $l = 1, \ldots, d$, $k_l(\cdot, \cdot)$ is the reproducing kernel of a RKHS $\mathcal{F}_l$ of real functions depending only on variable x_l and such that $1 \notin \mathcal{F}_l$. **Without constant functions**

In addition, for all $l = 1, \ldots, d$ and $\forall x_l \in \mathcal{X}_l$, we have

Zero-mean kernel

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0. \tag{11}$$

Kernel-embedding of probability distributions for GSA: HSIC

New HSIC-based sensitivity index

- Also a decomposition
- Can handle easily structured outputs
- **Generalization of MMD-based index!**

Kernel more or less converging to a dirac

Proposition 2. *For all subset $A \subseteq \mathcal{P}_d$, let us define a product RKHS $\mathcal{H}_A = \mathcal{F}_A \times \mathcal{G}$ with kernel $k_A(\mathbf{x}_A, \mathbf{x}'_A) k_\mathcal{Y}(y, y')$. We further assume that $\forall \mathbf{x}_A \in \mathcal{X}_A$, $p_{\mathbf{X}_A}(\mathbf{x}_A) > 0$ and that*

$$k_A(\mathbf{x}_A, \mathbf{x}'_A) = \frac{1}{\sqrt{p_{\mathbf{X}_A}(\mathbf{x}_A)}\sqrt{p_{\mathbf{X}_A}(\mathbf{x}'_A)}} \prod_{l \in A} \frac{1}{h} K\left(\frac{x_l - x'_l}{h}\right) \tag{13}$$

where $K : \mathbb{R} \to \mathbb{R}$ is a symmetric kernel function satisfying $\int_u K(u)du = 1$, and $h > 0$. Then we have $\forall A \subseteq \mathcal{P}_d$

$$\lim_{h \to 0} \mathrm{HSIC}(\mathbf{X}_A, Y) = \mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right)$$

where $\mathrm{HSIC}(\mathbf{X}_A, Y)$ is defined with the product RKHS $\mathcal{H}_A = \mathcal{F}_A \times \mathcal{G}$ and $\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})$ with the RKHS $\mathcal{G}$.

Kernel-embedding of probability distributions for GSA: HSIC

Wait a minute!

In addition, for all $l = 1, \ldots, d$ *and* $\forall x_l \in \mathcal{X}_l$*, we have*

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0. \quad (11)$$

Zero-mean kernel

> **How do we build a kernel satisfying this?**

Kernel-embedding of probability distributions for GSA: HSIC

Zero-mean kernel

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0.$$

Easy case: inputs are uniform on [0,1]

- We can directly use famous Sobolev kernels (from SS-ANOVA, COSSO, ACOSSO, …)

$$k_l(x_l, x_l') = \frac{B_{2r}(|x_l - x_l'|)}{(-1)^{r+1}(2r)!} + \sum_{j=1}^{r} \frac{B_j(x_l) B_j(x_l')}{(j!)^2}$$

where B are Bernoulli polynomials.

- Always possible to transform independent inputs to end up with this case (via probability integral transform)
- But sensitivity index is not invariant via nonlinear transformations

Kernel-embedding of probability distributions for GSA: HSIC

Zero-mean kernel

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0.$$

General case 1

- Kernels built by Durrande et al. (2012) in the context of GP models with ANOVA decomposition inside

$$k_0^D(x, x') = k(x, x') - \frac{\int k(x,t) d\mathrm{P}(t) \int k(x',t) d\mathrm{P}(t)}{\iint k(s,t) d\mathrm{P}(s) d\mathrm{P}(t)}$$

- Built from any initial kernel k
- Very nice theory, but needs numerical integration to compute the second term in general

Kernel-embedding of probability distributions for GSA: HSIC

Zero-mean kernel

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0.$$

General case 2

- Kernels introduced in the context of Stein discrepancy in lieu of MMD

$$k_0^S(\mathbf{x},\mathbf{x}') = \nabla_{\mathbf{x}}\nabla_{\mathbf{x}'}k(\mathbf{x},\mathbf{x}') + \frac{\nabla_{\mathbf{x}}p(\mathbf{x})}{p(\mathbf{x})}\nabla_{\mathbf{x}'}k(\mathbf{x},\mathbf{x}') + \frac{\nabla_{\mathbf{x}'}p(\mathbf{x}')}{p(\mathbf{x}')}\nabla_{\mathbf{x}}k(\mathbf{x},\mathbf{x}') + \frac{\nabla_{\mathbf{x}}p(\mathbf{x})}{p(\mathbf{x})}\frac{\nabla_{\mathbf{x}'}p(\mathbf{x}')}{p(\mathbf{x}')}k(\mathbf{x},\mathbf{x}')$$

- Built from any initial kernel k again, but must be differentiable this time
- Needs derivative of the log pdf of the inputs
- Means that we only need to know the pdf up to a constant
  - **A potential interest for GSA problems where some inputs are obtained through Bayesian calibration**

Proof outline for ANOVA decomposition of HSIC (1/2)

First assume that Mercer's theorem holds $k_{\mathcal{Y}}(y, y') = \sum_{r=1}^{\infty} \phi_r(y)\phi_r(y')$

Then write HSIC as

$$\mathrm{HSIC}(\mathbf{X}, Y) = \sum_{r=1}^{\infty} \|g^{[r]}\|_{\mathcal{F}}^2 \qquad g^{[r]}(\mathbf{x}) = \int_{\mathcal{X}} \int_{\mathcal{Y}} k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}')\phi_r(y') \left[p_{\mathbf{X}Y}(\mathbf{x}', y') - p_{\mathbf{X}}(\mathbf{x}')p_Y(y')\right] d\mathbf{x}' dy'$$

Key part: orthogonal decomposition of each g function thanks to Kuo et al. (2010)

- This is where we need the strong assumptions on the input kernels

$$g^{[r]} = \sum_{A \subseteq \mathcal{P}_d} g_A^{[r]}$$

$$g_A^{[r]} = \sum_{B \subseteq A} (-1)^{|A|-|B|} P_{-B}(g^{[r]})$$

Proof outline for ANOVA decomposition of HSIC (2/2)

We then plug the decompositions inside HSIC

$$
\begin{aligned}
\mathrm{HSIC}(\mathbf{X}, Y) &= \sum_{r=1}^{\infty} \|g^{[r]}\|_{\mathcal{F}}^2 \\
&= \sum_{A \subseteq \mathcal{P}_d} \sum_{r=1}^{\infty} \|g_A^{[r]}\|_{\mathcal{F}}^2 \\
&= \sum_{A \subseteq \mathcal{P}_d} \sum_{B \subseteq A} (-1)^{|A|-|B|} \sum_{r=1}^{\infty} \|P_{-B}(g^{[r]})\|_{\mathcal{F}}^2
\end{aligned}
$$

And the final result comes from rewriting the projections

$$
\begin{aligned}
\sum_{r=1}^{\infty} \|P_{-B}(g^{[r]})\|_{\mathcal{F}}^2 &= \sum_{r=1}^{\infty} \int_{\mathcal{X}_B \times \mathcal{X}_B} \int_{\mathcal{Y} \times \mathcal{Y}} k_B(\mathbf{x}_B, \mathbf{x}'_B) \phi_r(y) \phi_r(y') \left[p_{\mathbf{X}_B Y}(\mathbf{x}_B, y) - p_{\mathbf{X}_B}(\mathbf{x}_B) p_Y(y)\right] \\
&\quad \left[p_{\mathbf{X}_B Y}(\mathbf{x}'_B, y') - p_{\mathbf{X}_B}(\mathbf{x}'_B) p_Y(y')\right] d\mathbf{x}_B d\mathbf{x}'_B dy dy' \\
&= \int_{\mathcal{X}_B \times \mathcal{X}_B} \int_{\mathcal{Y} \times \mathcal{Y}} k_B(\mathbf{x}_B, \mathbf{x}'_B) \left(\sum_{r=1}^{\infty} \phi_r(y) \phi_r(y') \right) \left[p_{\mathbf{X}_B Y}(\mathbf{x}_B, y) - p_{\mathbf{X}_B}(\mathbf{x}_B) p_Y(y)\right] \\
&\quad \left[p_{\mathbf{X}_B Y}(\mathbf{x}'_B, y') - p_{\mathbf{X}_B}(\mathbf{x}'_B) p_Y(y')\right] d\mathbf{x}_B d\mathbf{x}'_B dy dy' \\
&= \int_{\mathcal{X}_B \times \mathcal{X}_B} \int_{\mathcal{Y} \times \mathcal{Y}} k_B(\mathbf{x}_B, \mathbf{x}'_B) k_{\mathcal{Y}}(y, y') \left[p_{\mathbf{X}_B Y}(\mathbf{x}_B, y) - p_{\mathbf{X}_B}(\mathbf{x}_B) p_Y(y)\right] \\
&\quad \left[p_{\mathbf{X}_B Y}(\mathbf{x}'_B, y') - p_{\mathbf{X}_B}(\mathbf{x}'_B) p_Y(y')\right] d\mathbf{x}_B d\mathbf{x}'_B dy dy' \\
&= \mathrm{HSIC}(\mathbf{X}_B, Y).
\end{aligned}
$$