Kernel-based Sensitivity Analysis of set-valued models. Application to pollutant concentration maps SA

Noé Fellmann
Céline Helbert & Christophette Blanchet (ECL)
Adrien Spagnol & Delphine Sinoquet (IFP Energies Nouvelles)

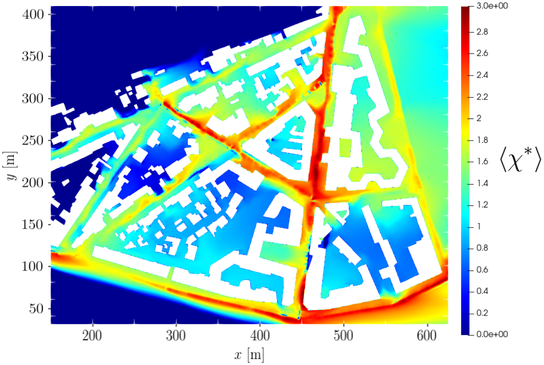
MASCOTNUM 2024

Model of pollutant concentration maps [Pasquier et al. 2023]

Input $\boldsymbol{U} = (\theta, U_\infty, q, \beta, \nu_{max})$ with,

- Wind direction $\theta \sim \mathcal{N}_{[5,10]}(0.7, 0.5^2)$ [rad]
- Wind speed $U_\infty \sim \mathcal{N}_{[-0.35;1.75]}(8, 2)$ [m/s]
- Traffic volume $q \sim \mathcal{SN}_{[100;500]}(450, 100, -3)$ [vehicle/h]
- Proportion of diesel and petrol engine $\beta \sim \mathcal{U}([0, 1])$ [-]
- Speed limit $\nu_{\max} \sim \mathcal{U}([30; 50])$ [km/h]

Output : $(x,y) \mapsto \Phi_{\boldsymbol{U}}(x, y)$ a pollutant concentration map



Sensitivity Analysis of map-valued models

Map-valued model :

$$\Phi : \boldsymbol{U} \mapsto \Phi_{\boldsymbol{U}}$$

where

$$\Phi_{\boldsymbol{U}} : (x,y) \mapsto \Phi_{\boldsymbol{U}}(x, y) \in \mathbb{R}$$

Goal : Quantify the effect of the inputs $\boldsymbol{U}$ on the spatial output $\Phi_{\boldsymbol{U}}$

Sensitivity Analysis of map-valued models

Map-valued model :

$$\Phi : \boldsymbol{U} \mapsto \Phi_{\boldsymbol{U}}$$

where

$$\Phi_{\boldsymbol{U}} : (x,y) \mapsto \Phi_{\boldsymbol{U}}(x,y) \in \mathbb{R}$$

Goal : Quantify the effect of the inputs $\boldsymbol{U}$ on the spatial output $\Phi_{\boldsymbol{U}}$

Sensitivity analysis context

$$(U_1, \ldots, U_d) \overset{f}{\mapsto} Z = f(U_1, \ldots, U_d)$$

How can the uncertainty of Z be divided and allocated to the uncertainty of the inputs U_i ?

- Sobol indices : $S_i = \frac{\mathrm{Var}\,\mathbb{E}(Z|U_i)}{\mathrm{Var}\,Z}$
- Dependence measures : $S_i = ||\mathbb{P}_{(U_i,Z)} - \mathbb{P}_{U_i} \otimes \mathbb{P}_Z||$

Screening : $U_1, \ldots, U_k$ are influential and $U_{k+1}, \ldots U_d$ are not influential
Ranking : $U_1 \prec \ldots \prec U_d$

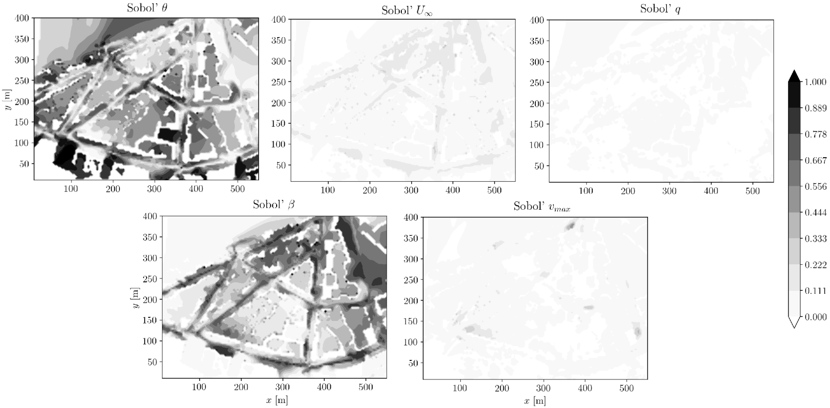
Table of Contents

Table of Contents

Sobol indices maps [Pasquier et al. 2023]

Sobol indices at the position (x, y) :

$$S_i(x, y) = \frac{\mathrm{Var}[\mathbb{E}[\Phi_{\boldsymbol{U}}(x, y)|U_i]]}{\mathrm{Var}[\Phi_{\boldsymbol{U}}(x, y)]}.$$



Aggregated Sobol indices [Pasquier et al. 2023]

Aggregated Sobol' indices (Gamboa, Janon et al. 2013) :

$$S_i^{\text{gen}} := \sum_{j=1}^{m} w_j S_i(x^{(j)}, y^{(j)}) \quad \text{with} \qquad w_j = \frac{\mathrm{Var}[\Phi_{\boldsymbol{U}}(x^{(j)}, y^{(j)})]}{\sum_{k=1}^{m} \mathrm{Var}[\Phi_{\boldsymbol{U}}(x^{(k)}, y^{(k)})]}.$$

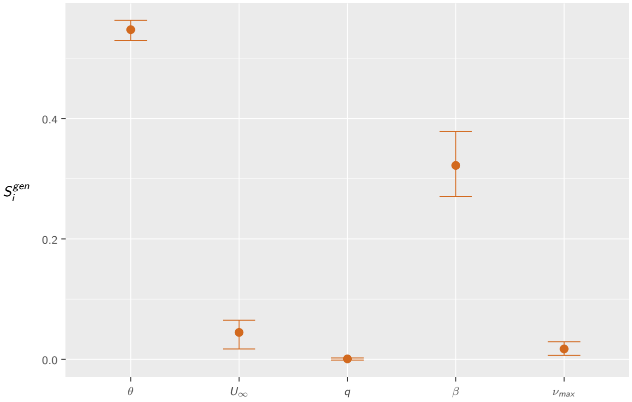
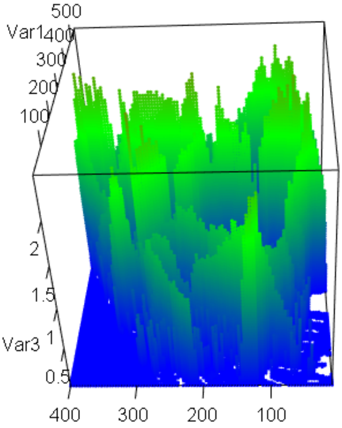


Figure – Estimated aggregated first-order Sobol' indices with 2^{12} model evaluations.

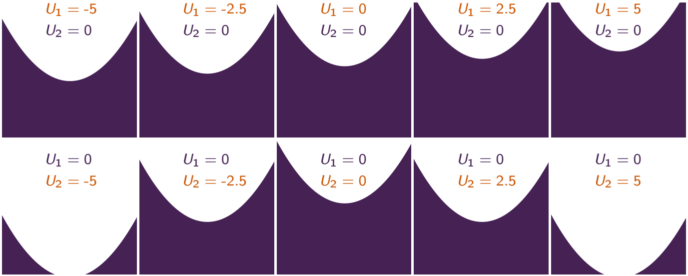
Table of Contents

From map-valued model to set-valued model

$$\Psi : \begin{array}{ccc} \mathcal{U} & \longrightarrow & \mathcal{L}(\mathcal{X}) \\ \boldsymbol{u} & \mapsto & \Gamma_{\boldsymbol{u}} = \{(x, y, c) \in \mathcal{D} \times [0, C_{max}],\ c \leq \Phi_{\boldsymbol{u}}(x, y)\} \end{array}$$



Sensitivity analysis of set-valued models ?



How to do sensitivity analysis of set-valued models ?

Sensitivity analysis on the volume

How to do sensitivity analysis of set-valued models ?

- Conduct sensitivity analysis on the volume : $U \to \lambda(\Gamma_U)$

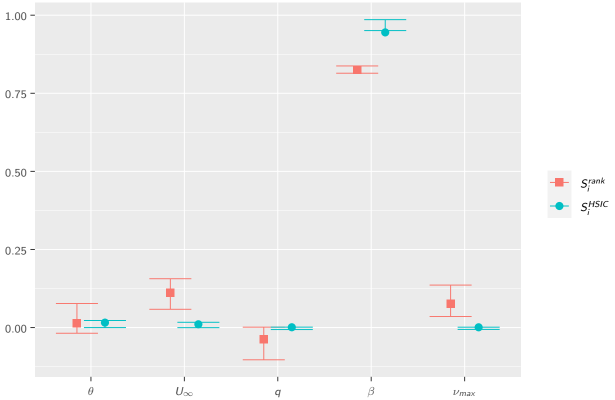


Figure – Estimation of Sobol indices (rank-method) and HSIC-based indices (rbf_anova kernel) of the volume of Γ_U. 1000 model evaluations are used and confidence interval are estimated with 100 bootstrap resamples.

Sensitivity analysis on the volume

How to do sensitivity analysis of set-valued models ?

- ~~Conduct sensitivity analysis on the volume : $U \to \lambda(\Gamma_U)$~~

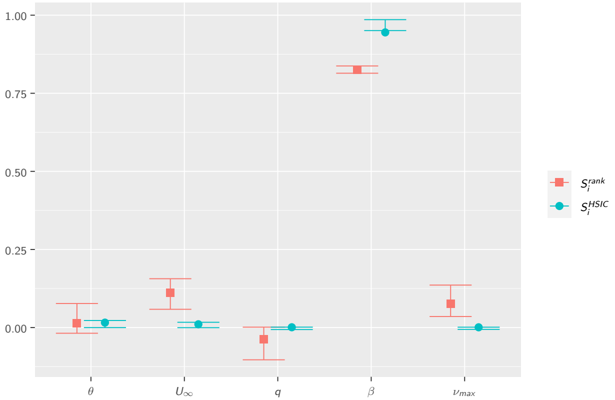


Figure – Estimation of Sobol indices (rank-method) and HSIC-based indices (rbf_anova kernel) of the volume of Γ_U. 1000 model evaluations are used and confidence interval are estimated with 100 bootstrap resamples.

Table of Contents

Distribution embedding into a RKHS

Given a kernel $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$,

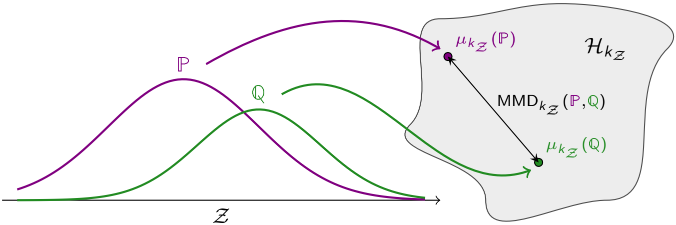


Figure – Kernel mean embedding

- $\mu_{k_{\mathcal{Z}}}(\mathbb{P}) = \int_{\mathcal{Z}} k(\cdot, z) d\mathbb{P}(z)$
- $\mathrm{MMD}_{k_{\mathcal{Z}}}(\mathbb{P}, \mathbb{Q}) = ||\mu_{k_{\mathcal{Z}}}(\mathbb{P}) - \mu_{k_{\mathcal{Z}}}(\mathbb{Q})||_{\mathcal{H}_{k_{\mathcal{Z}}}}$

The MMD is a distance between distribution iif $k_{\mathcal{Z}}$ is a characteristic kernel, i.e. the mean embedding is injective.

From MMD to HSIC

Dependence measure $S_i = ||\mathbb{P}_{(U_i,Z)} - \mathbb{P}_{U_i} \otimes \mathbb{P}_Z||$ to quantify the effect of an input U_i on the entire output distribution.

Hilbert Schmidt Independence Criterion (HSIC), Gretton et al. 2006

With $K_i = k_{\mathcal{U}_i} \otimes k_{\mathcal{Z}}$, the HSIC is given by :

$$\begin{aligned}\mathrm{HSIC}_{K_i}(U_i, Z) :&= \mathrm{MMD}^2_{K_i}(\mathbb{P}_{U_i,Z}, \mathbb{P}_{U_i} \otimes \mathbb{P}_Z) \\ &= ||\mu_{K_i}(\mathbb{P}_{U_i,Z}) - \mu_{k_{\mathcal{U}_i}}(\mathbb{P}_{U_i}) \otimes \mu_{k_{\mathcal{Z}}}(\mathbb{P}_Z)||^2_{\mathcal{H}_K}\end{aligned}$$

If $k_{\mathcal{U}_i}$ and $k_{\mathcal{Z}}$ are characteristic, then

$$\mathrm{HSIC}_{K_i}(U_i, Z) = 0 \text{ iif } U_i \perp Z$$

Screening is then possible with independence testing.

HSIC-ANOVA indices [daVeiga 2021]

Assuming that the inputs are independent and that the input kernels are ANOVA,

$$\mathrm{HSIC}(\boldsymbol{U}, Z) = \sum_{A \subseteq \{1,\dots,d\}} \sum_{B \subseteq A} (-1)^{|A|-|B|} \, \mathrm{HSIC}\left(\boldsymbol{U}_B, Z\right).$$

HSIC-ANOVA indices are then defined as :

$$S_i^{\mathrm{HSIC}} := \frac{\mathrm{HSIC}(U_i, Z)}{\mathrm{HSIC}(\boldsymbol{U}, Z)},$$

$$S_{T_i}^{\mathrm{HSIC}} := 1 - \frac{\mathrm{HSIC}(\boldsymbol{U}_{-i}, Z)}{\mathrm{HSIC}(\boldsymbol{U}, Z)}$$

and are suited for ranking (and screening).

- Easy to estimate :

$$\mathrm{HSIC}(\boldsymbol{U}_A, Z) = \mathbb{E}\left[(K_A(\boldsymbol{U}_A, \boldsymbol{U}_A{}') - 1) k_{\mathcal{Z}}(Z, Z')\right].$$

- Only requirement : to have kernels on the inputs and on the output

Table of Contents

HSIC ANOVA indices for sets, definition of the indices

$\mathcal{Z}$	$\longleftrightarrow$	$\mathscr{L}(\mathcal{X})$ the space of Lebesgue measurable subsets of $\mathcal{X}$
Z	$\longleftrightarrow$	Γ a random set
$k_{\mathcal{Z}}$	$\longleftrightarrow$	k_{set} a (characteristic) kernel on $\mathscr{L}(\mathcal{X})$

Proposition (Fellmann, Blanchet-Scalliet et al. 2023)

The function k_{set} defined by :

$$k_{set}(\gamma_1, \gamma_2) = \exp\left(-\frac{\lambda(\gamma_1 \Delta \gamma_2)}{2\sigma^2}\right)$$

- *is a bounded and measurable kernel*
- *is characteristic.*

$$\mathrm{HSIC}_{k_i \otimes k_{set}}(U_i, \Gamma) = ||\mu_{k_i \otimes k_{set}}(\mathbb{P}_{(U_i,\Gamma)}) - \mu_{k_i \otimes k_{set}}(\mathbb{P}_{U_i} \otimes \mathbb{P}_{\Gamma})||^2_{\mathcal{H}_{k_i \otimes k_{set}}}$$

HSIC ANOVA indices for sets, estimation

$$\mathrm{HSIC}(\boldsymbol{U}_A, \Gamma) = \mathbb{E}\left[(K_A(\boldsymbol{U}_A, \boldsymbol{U}_A{}') - 1)k_{set}(\Gamma, \Gamma')\right].$$

The indices can be estimated using U-statistics :

$$\widehat{\mathrm{H}_{set}}\,(U_i, \Gamma) = \frac{2}{n(n-1)} \sum_{j<l}^{n} \left(K_i\left(U_i^{(j)}, U_i^{(l)}\right) - 1\right) \exp(-\frac{\lambda(\mathcal{X})}{2\sigma^2} \hat{\lambda}_m(\Gamma^{(j)} \Delta \Gamma^{(l)})).$$

Input kernels :

- the Sobolev kernel of order 1, $k_{sob}(x, y) = 1 + (x - \frac{1}{2})(y - \frac{1}{2}) + \frac{1}{2}[(x-y)^2 - |x-y| + \frac{1}{6}]$
- the Gaussian kernel, $k_{rbf}(x, y) = e^{-\frac{1}{2}\left(\frac{x-y}{\sigma}\right)^2}$ with $\sigma > 0$,
- the Laplace kernel, $k_{exp}(x, y) = e^{-\frac{|x-y|}{h}}$ with $h > 0$,
- the Matérn 3/2, $k_{3/2}(x, y) = \left(1 + \sqrt{3}\frac{|x-y|}{h}\right) e^{-\sqrt{3}\frac{|x-y|}{h}}$ with $h > 0$,
- the Matérn 5/2, $k_{5/2}(x, y) = \left(1 + \sqrt{5}\frac{|x-y|}{h} + \frac{5}{3}\frac{|x-y|}{h^2}\right) e^{-\sqrt{5}\frac{|x-y|}{h}}$ with $h > 0$.

HSIC ANOVA indices for sets, results

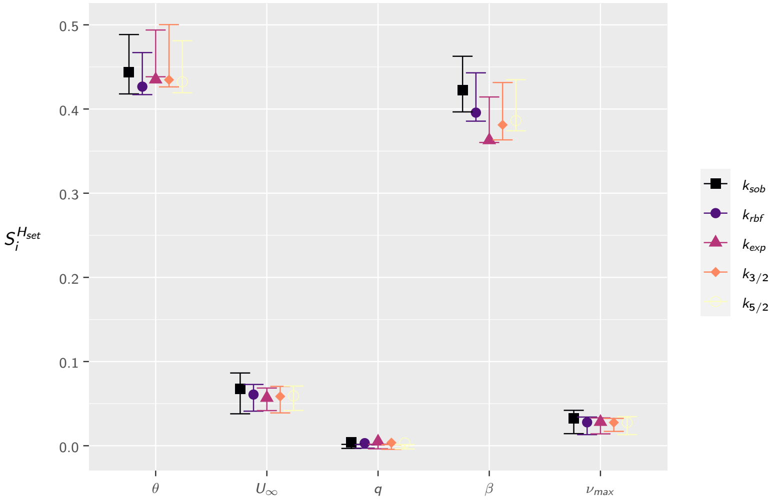


Figure – Estimation of $S_i^{H_{set}}$ for five input kernels, 1000 model evaluations. Confidence intervals are obtained by bootstrap with 100 resamples

Comparison with other indices

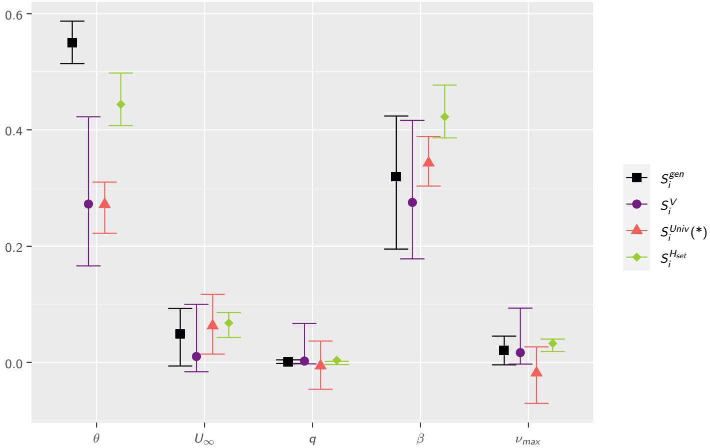


Figure – Comparison of the four indices with a total budget of $n = 1000$ model evaluations. 100 bootstrap sample are used to estimate confidence intervals [Fellmann, Pasquier et al. 2023]

* Gamboa, Klein et al. 2021

Comparison with other indices

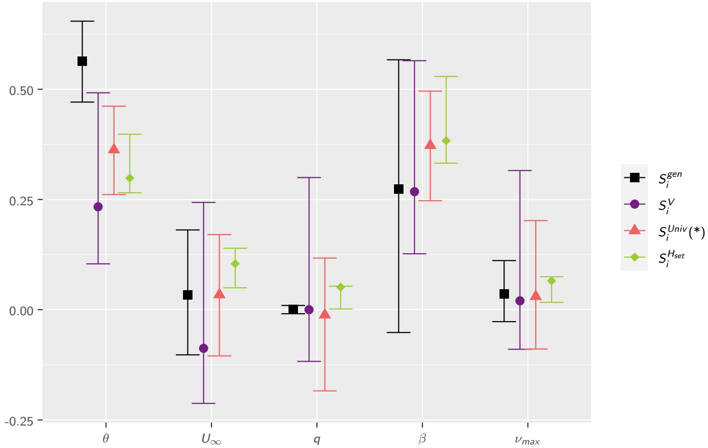


Figure – Comparison of the four indices with a total budget of $n = 100$ model evaluations. 100 bootstrap sample are used to estimate confidence intervals [Fellmann, Pasquier et al. 2023]

* Gamboa, Klein et al. 2021

Conclusion

- HSIC-ANOVA indices are
  - ✓ cheap to estimate
  - ✓ suited for screening
  - ✓ suited for ranking
  - ✗ sometimes hard to interpret (interactions)
  - ✓ very permissive in the type of the output (and of the inputs ?)
- The presented methodology for map-valued outputs but can be used for any set-valued outputs.
- Current work : SA of the excursion sets of a robust optimization problem

Conclusion

- HSIC-ANOVA indices are
  - ✓ cheap to estimate
  - ✓ suited for screening
  - ✓ suited for ranking
  - ✗ sometimes hard to interpret (interactions)
  - ✓ very permissive in the type of the output (and of the inputs ?)
- The presented methodology for map-valued outputs but can be used for any set-valued outputs.
- Current work : SA of the excursion sets of a robust optimization problem
- Thanks !

daVeiga, Sébastien (jan. 2021). "Kernel-based ANOVA decomposition and Shapley effects - Application to global sensitivity analysis". working paper or preprint. url : https://hal.archives-ouvertes.fr/hal-03108628.

Fellmann, Noé, Christophette Blanchet-Scalliet et al. (2023). Kernel-based sensitivity analysis for (excursion) sets. arXiv : 2305.09268 [math.ST].

Fellmann, Noé, Mathis Pasquier et al. (2023). "Sensitivity analysis for sets : application to pollutant concentration maps". In : arXiv preprint.

Gamboa, Fabrice, Alexandre Janon et al. (2013). Sensitivity analysis for multidimensional and functional outputs. arXiv : 1311.1797 [stat.AP].

Gamboa, Fabrice, Thierry Klein et al. (août 2021). "Sensitivity analysis in general metric spaces". In : Reliability Engineering and System Safety. doi : 10.1016/j.ress.2021.107611. url : https://hal.archives-ouvertes.fr/hal-02044223.

Gretton, Arthur et al. (2006). "A Kernel Method for the Two-Sample-Problem". In : Advances in Neural Information Processing Systems. T. 19. MIT Press. url : https://proceedings.neurips.cc/paper/2006/hash/e9fb2eda3d9c55a0d89c98d6c54b5b3e-Abstract.html.

Pasquier, Mathis et al. (2023). "A Lattice-Boltzmann-based modelling chain for traffic-related atmospheric pollutant dispersion at the local urban scale". In : Building and Environment 242, p. 110562.

HSIC (no ANOVA) indices for sets, results

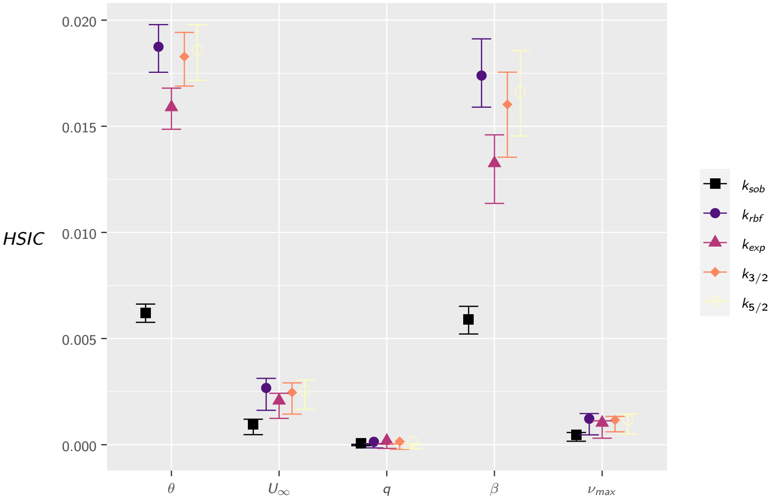


Figure – Estimation of *HSIC* for five input kernels, 1000 model evaluations. Confidence intervals are obtained by bootstrap with 100 resamples

HSIC ANOVA indices for sets, results

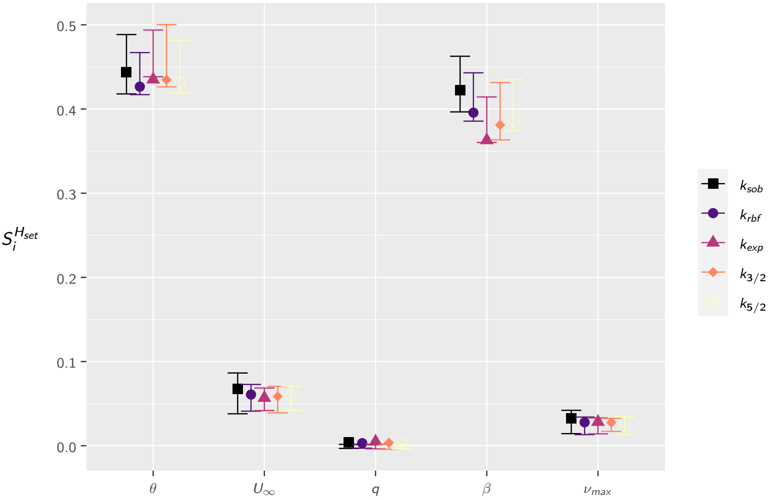


Figure – Estimation of $S_i^{\mathrm{H}set}$ for five input kernels, 1000 model evaluations. Confidence intervals are obtained by bootstrap with 100 resamples