SENSITIVITY ANALYSIS: PAWN SENSITIVITY INDEX

Course name: Advanced Probability and Statistic

Instructor: Dr. Mohsen Nasseri, Associate Prof., School of Civil Engineering, University of Tehran.

Students: Fatemeh Mikaeli, Student Number:810201140

Nastaran Onsori, Student Number:810201103

Farnaz Sadat Shahi, Student Number: 810201086

Project: Sensitivity Analysis: PAWN Sensitivity Index

Contents

Preface

Sensitivity analysis:

Sensitivity analysis is the study of how the uncertainty in the output of a mathematical model or system (numerical or otherwise) can be divided and allocated to different sources of uncertainty in its inputs. A mathematical model (for example in biology, climate change, economics or engineering) can be highly complex, and as a result, its relationships between inputs and outputs may be poorly understood. In such cases, the model can be viewed as a black box, i.e. the output is an "opaque" function of its inputs. Quite often, some or all of the model inputs are subject to sources of uncertainty, including errors of measurement, absence of information and poor or partial understanding of the driving forces and mechanisms. This uncertainty imposes a limit on our confidence in the response or output of the model. Further, models may have to cope with the natural intrinsic variability of the system (aleatory), such as the occurrence of stochastic events. Good modeling practice requires that the modeler provide an evaluation of the confidence in the model. This requires, first, a quantification of the uncertainty in any model results (uncertainty analysis); and second, an evaluation of how much each input is contributing to the output uncertainty. Sensitivity analysis addresses the second of these issues (although uncertainty analysis is usually a necessary precursor), performing the role of ordering by importance the strength and relevance of the inputs in determining the variation in the output.

Sensitivity analysis methods

One-at-a-time (OAT): One of the simplest and most common approaches is that of changing one-factor-at-a-time (OAT), to see what effect this produces on the output.

Derivative-based local methods: Local derivative-based methods involve taking the partial derivative of the output Y with respect to an input factor Xi.

Regression analysis: Regression analysis, in the context of sensitivity analysis, involves fitting a linear regression to the model response and using standardized regression coefficients as direct measures of sensitivity. The regression is required to be linear with respect to the data.

Variance-based methods: Variance-based methods are a class of probabilistic approaches which quantify the input and output uncertainties as probability distributions, and decompose the output variance into parts attributable to input variables and combinations of variables. The sensitivity of the output to an input variable is therefore measured by the amount of variance in the output caused by that input.

Variogram analysis of response surfaces (VARS): One of the major shortcomings of the previous sensitivity analysis methods is that none of them considers the spatially ordered structure of the response surface/output of the model Y=f(X) in the parameter space. By utilizing the concepts of directional variograms and covariograms, variogram analysis of response surfaces (VARS) addresses this weakness through recognizing a spatially continuous correlation structure to the values of Y, and hence also to the values of $\frac{\delta Y}{\delta xi}$.

Screening: Screening is a particular instance of a sampling-based method. The objective here is rather to identify which input variables are contributing significantly to the output uncertainty in high-dimensionality models, rather than exactly quantifying sensitivity (i.e. in terms of variance).

Scatter plots: A simple but useful tool is to plot scatter plots of the output variable against individual input variables, after (randomly) sampling the model over its input distributions.

Local and global sensitivity analysis:

Historically, there have been two broad categories of sensitivity analysis techniques: local and global. Local sensitivity analysis is performed by varying model parameters around specific reference values, with the goal of exploring how small input perturbations influence model performance. Due to its ease-of-use and limited computational demands, this approach has been widely used in literature, but has important limitations. If the model is not linear, the results of local sensitivity analysis can be heavily biased, as they are strongly influenced by independence assumptions and a limited exploration of model inputs. Local sensitivity analysis considers the output variability against variations of an input factors around a specific value, while global sensitivity analysis (or GSA) considers variations within the entire space of variability of the input factors. The application of local SA obviously requires the user to specify a nominal value for the input factors.

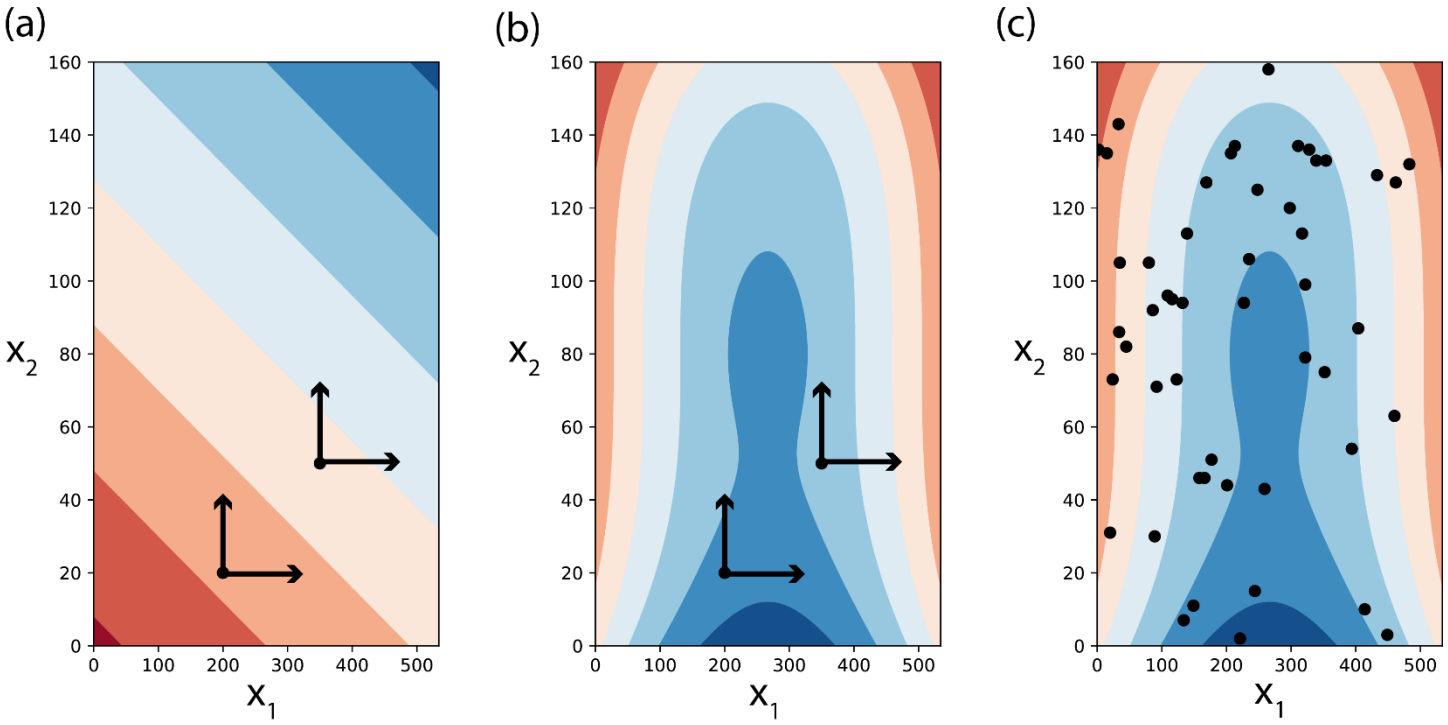


Figure 1: Treatment of a two-dimensional space of variability by local (panels a-b) and global (panel c) sensitivity analyses.

In general, as local sensitivity analysis only partially and locally explores a model's parametric space, it is not considered a valid approach for nonlinear models. This is illustrated in Fig. 1 (a-b), presenting contour plots of a model response (y) with an additive linear model (a) and with a nonlinear model (b). In a linear model without interactions between the input terms, local sensitivity analysis (assuming deviations from some reference values) can produce appropriate sensitivity indices (Fig. 1 (a)). If however, factors interact, the local and partial consideration of the space can not properly account for each factor's effects on the model response (Fig. 1 (b)), as it is only informative at the reference value where it is applied. In contrast, a global sensitivity analysis varies uncertain factors within the entire feasible space of variable model responses (Fig. 1 (c)). This approach reveals the global effects of each parameter on the model output, including any interactive effects. For models that cannot be proven linear, global sensitivity analysis is preferred.

Purposes of sensitivity analysis:

Often, three specific purposes for global SA are defined:

- Factor Priorization (FP) aims at ranking the inputs xi in terms of their relative contribution to output uncertainty.
- Factor Fixing (FF), or screening, aims at determining the inputs, if any, that do not give any contribution to output uncertainty.
- Factor Mapping (FM) aims at determining the regions in the inputs space that produce specific output values, for instance above a prescribed threshold.

Global SA aimed at FP and FF often employees sensitivity indices, or importance measures. These are synthetic indices that quantify the relative contribution to output uncertainty from each input. A sensitivity index of zero means that the associated input is non-influential (which is useful for FF) while the higher the index the more influential the input (FP).

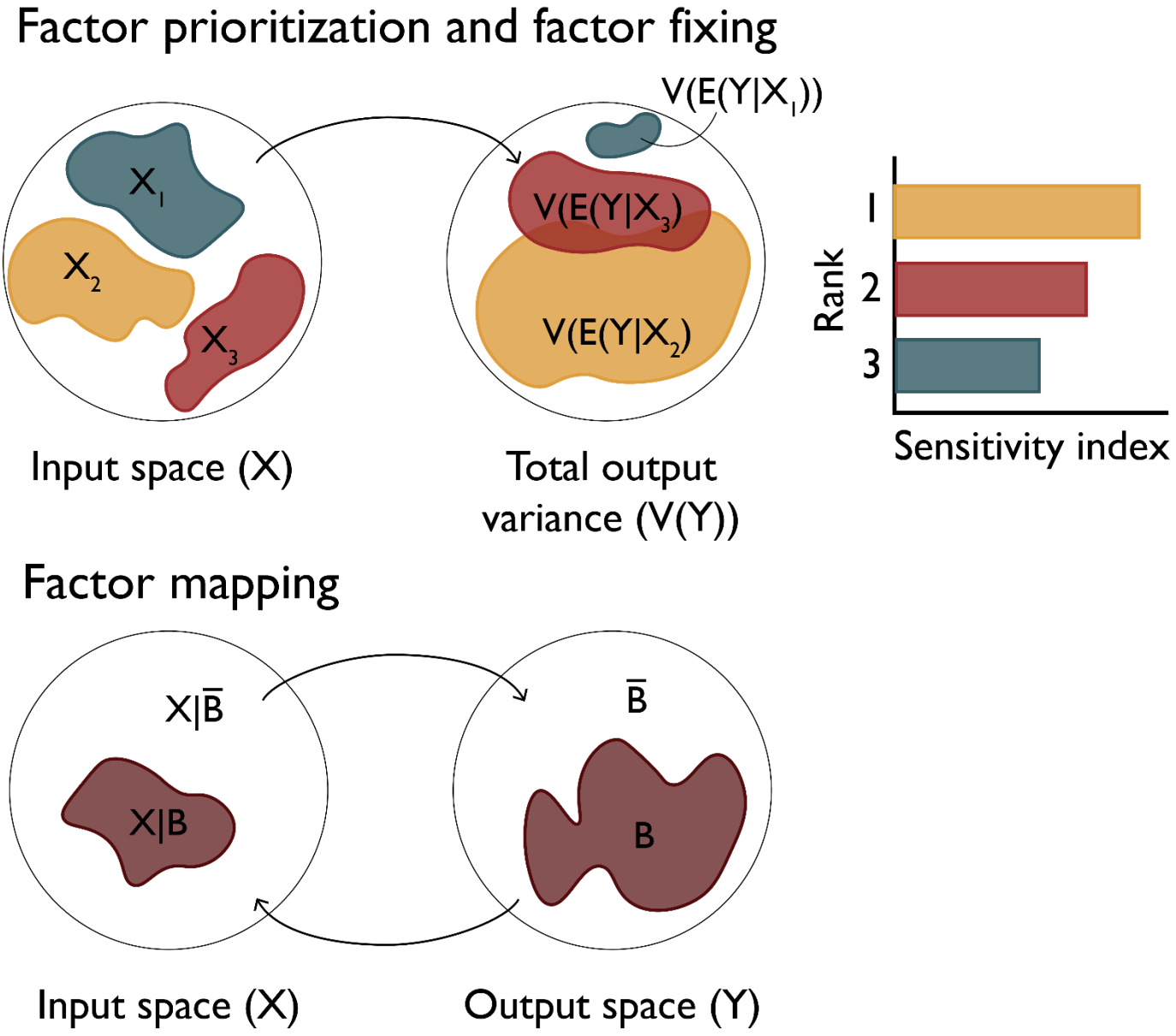


Figure 2: Factor prioritization, factor fixing and factor mapping settings of sensitivity analysis.

Variance-based Methods:

Variance-based sensitivity analysis methods hypothesize that various specified model factors contribute differently to the variation of model outputs; therefore, decomposition and analysis of output variance can determine a model's sensitivity to input parameters. The most popular variance-based method is the Sobol method, which is a global sensitivity analysis method that takes into account complex and nonlinear factor interaction when calculating sensitivity indices, and employs more sophisticated sampling methods (e.g., the Sobol sampling method). The Sobol method is able to calculate three types of sensitivity indices that provide different types of information about model sensitivities. These indices include first-order, higher-order (e.g., second-, third-, etc. orders), and total-order sensitivities.

The first-order sensitivity index indicates the percent of model output variance contributed by a factor individually (i.e., the effect of varying alone) and is obtained using the following:

$$Si^1 = \frac{Vi}{V(Y)} \qquad Vi = Var_{xi}(E_{X\sim i}(Y|Xi))$$

The X~i notation indicates the set of all variables except Xi.

Higher-order sensitivity indices explore the interaction between two or more parameters that contribute to model output variations

The total sensitivity analysis index represents the entire influence of an input factor on model outputs including all of its interactions with other factors. In other words, total order indices include first-order and all higher-order interactions associated with each factor and can be estimated calculated using the following:

$$STi = 1 - \frac{Var_{x\sim i}(E_{Xi}(Y|X\sim i))}{Var(Y)}$$

Besides the Sobol method, there are some other variance-based sensitivity analysis methods, such as the Fourier amplitude sensitivity test and extended-FAST, that have been used by the scientific community. However, Sobol remains by far the most common method of this class. Variance-based techniques have been widely used and have proved to be powerful in a variety of applications. Despite their popularity, some authors have expressed concerns about the methods' appropriateness in some settings. Specifically, the presence of heavy-tailed distributions or outliers, or when model outputs are multimodal can bias the sensitivity indices produced by these methods. Moment-independent measures attempt to overcome these challenges.

Moment-Independent (Density-Based) Methods:

These methods typically compare the entire distribution (i.e., not just the variance) of input and output parameters in order to determine the sensitivity of the output to a particular input variable. Several moment-independent sensitivity analysis methods have been proposed in recent years. The delta(δ) moment-independent method calculates the difference between unconditional and conditional cumulative distribution functions of the output. The method has become widely used in various disciplines. The sensitivity index is defined as follows:

$$S_i = \delta_i = \frac{1}{2} E_{x_i} |f_y(y) - f_{y|x_i}(y)| dy$$

where is the probability density function of the entire model output y, and $f_{y|xi}(y)$is the conditional density of y , given that factor xi assumes a fixed value. The δi sensitivity indicator therefore represents the normalized expected shift in the distribution of y provoked by xi. Moment independent methods are advantageous in cases

where we are concerned about the entire distribution of events, such as when uncertain factors lead to more extreme events in a system. Further, they can be used with a pre-existing sample of data, without requiring a specific sampling scheme, unlike the previously reviewed methods. The sensitivity index does not include interactions between factors and it is therefore akin to the first order index produced by the Sobol method. Interactions between factors can still be estimated using this method, by conditioning the calculation on more than one uncertain factor being fixed.

Density based indices are relatively less easy to implement than variance-based ones, mainly because their computation requires the knowledge of many conditional PDFs. As PDFs are generally unknown, empirical PDFs must be used. The simplest approach to derive an empirical PDF is to use a histogram of the data sample, however the resulting shape can be significantly affected by the position of the first bin and the bin width, whose appropriate values may be difficult to determine. PDFs are better approximated using kernel density estimation (KDE) methods, since they only require to specify a single parameter, the bandwidth.

The PAWN sensitivity index:

Density-based sensitivity indices are moment-independent indices because, by definition, they consider the entire probability distribution of the output rather than one of its moments only. They measure sensitivity by estimating the variations that are induced in the output distribution when removing the uncertainty about one (or more) inputs. More specifically, the sensitivity to input xi is measured by the distance between the unconditional probability distribution of y that is obtained when all inputs vary simultaneously, and the conditional distributions that are obtained when varying all inputs but xi (i.e. xi is fixed at a nominal value xi).

In our approach, and in contrast to other density-based approaches, the conditional and unconditional distributions are characterized by their Cumulative Distribution Functions (CDFs) rather than their Probability Distribution Functions (PDFs). The reason for preferring CDFs is that they are much easier to approximate.

As a measure of distance between unconditional and conditional CDFs, it use the Kolmogorov-Smirnov statistic. With respect to other distance measures, this has several advantages. First, it varies between 0 and 1 regardless of the range of variation of the model output y, which ensures that our sensitivity index is an absolute measure. Secondly, when using this approach for Factor Fixing we can build on the statistical results of the two-sample Kolmogorov-Smirnov test to determine non-influential inputs at a given confidence level.

In the following, the unconditional cumulative distribution function of the output y is denoted by Fy(y), and the conditional cumulative distribution function when xi is fixed is denoted by $F_{y|xi}(y)$. Since $F_{y|xi}(y)$ accounts for what happens when the variability due to xi is removed, its distance from Fy(y) provides a measure of the effects of xi on y. The limit case is when $F_{Y|xi}(Y)$ coincides with Fy(y): it means that removing the uncertainty about xi does not affect the output distribution, and one can conclude that xi has no influence on y. If instead the distance between $F_{Y|xi}(Y)$ and Fy(y) increases, the influence of xi increases as well. As a measure of distance between unconditional and conditional CDFs, we use the Kolmogorov-Smirnov statistic (see Kolmogorov–Smirnov statistic):

$$KS(x_i) = \max_y \left| F_y(y) - F_{y|x_i}(y) \right|$$

As KS depends on the value at which we fix xi, the PAWN index Ti considers a statistic (e.g. the maximum or the median) over all possible values of xi:

$$T_i = \operatorname*{stat}_{x_i}[KS(x_i)]$$

By definition:

1. Ti varies between 0 and 1.
2. The lower the value of Ti, the less influential xi.
3. If Ti=0, then xi has no influence on y.

The PAWN index Ti can be used for both Factor Prioritization and Factor Fixing. It is global, quantitative and model independent. It is unconditional on any assumed input value because the dependency the KS statistic on the value of xi is removed by the statistic over xi that appears in Ti formula. It is easy to interpret because it is an absolute measure and thus its numerical value does not depend on the units of measurements of y. It is easy to compute because it can be easily approximated also when CDFs are not known.

One more property of index Ti is that it can be easily tailored to focus on a particular sub-range of the output distribution, rather than considering the entire range. To this end, it is sufficient that the maximum appearing in KS Equation be taken with respect to y values in the sub-range of interest.

Numerical implementation

First of all, the Kolmogorov-Smirnov statistic is approximated by:

$$\widehat{KS}(x_i) = \max_y \left| \widehat{F}_y(y) - \widehat{F}_{y|x_i}(y) \right|$$

where $\widehat{F}_y(.)$and $\widehat{F}y|xi(.)$are the empirical unconditional and conditional CDFs. The unconditional CDF $\widehat{F}_y(.)$ is approximated using Nu output evaluations obtained by sampling the entire input feasibility space. The conditional CDF $\widehat{F}y|xi(.)$ instead is approximated using Nc output evaluations obtained by sampling the non-fixed inputs only, while keeping the value of xi fixed.

Secondly, the statistic with respect to the conditioning value of xi is replaced by its sample version:

$$\widehat{T}_i = \underset{x_i = \bar{x}_i^{(1)}, \ldots, \bar{x}_i^{(n)}}{\text{stat}} \left[\left(\widehat{KS}(x_i) \right) \right]$$

The total number of model evaluations necessary to compute the sensitivity indices for all the M inputs is:

$$Nu + n * Nc * M$$

The values for n, Nu and Nc can be chosen by trial-and-error. As n sets the number of conditioning values sampled from the one dimensional space of variation of xi, its value might reasonably be in the order of few dozens (for example n between 10 and 50), while Nu and Nc, which set the number of samples taken in the M-dimensional and (M - 1)-dimensional space of all inputs and all-inputs-but-xi, should be significantly higher. Still, given the regularity properties of CDFs (continuity, monotonicity, relative smoothness) our approximation strategy is quite effective even when using small values for Nc and Nu (in the order of few hundreds), to limit the total number of model evaluations. Furthermore, given the simplicity of the computation of the PAWN index, its robustness to the selected values of Nc and Nu can be quickly estimated by repeating computations using bootstrap resamples of different sizes from the same dataset of input/output samples.

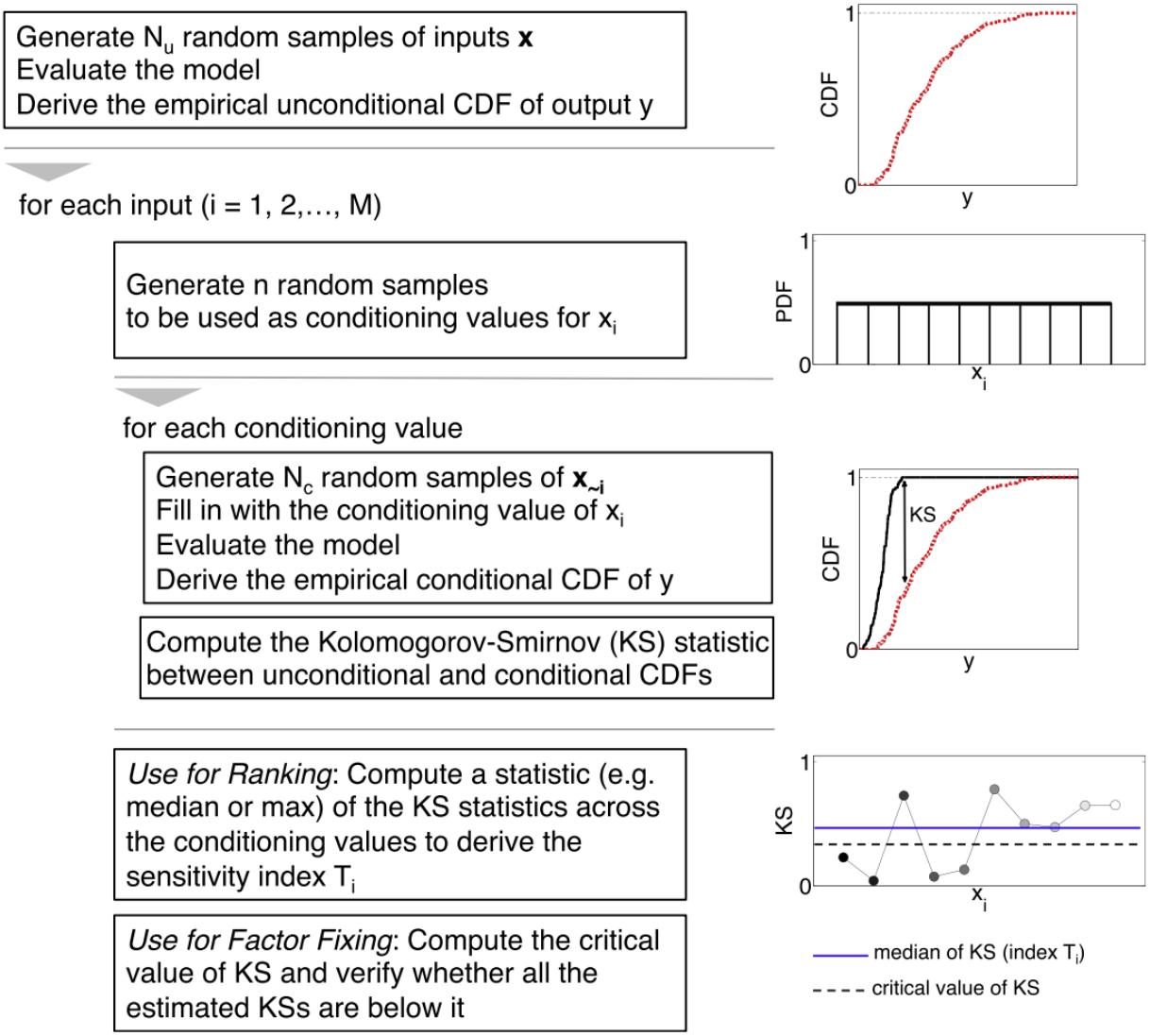


Figure 3: The steps in the numerical implementation of the PAWN index. Here, $x=|x_1,\ldots,x_M|$ *is the vector including all the uncertain inputs and* $x_{\sim i}=|x1,\ldots xi\text{-}1,xi+1,\ldots,xm|$ *is the vector of all the inputs but the i-th*

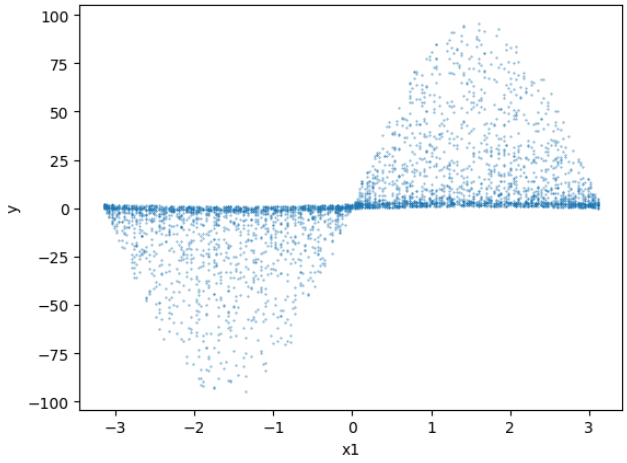
Numerical example 1:

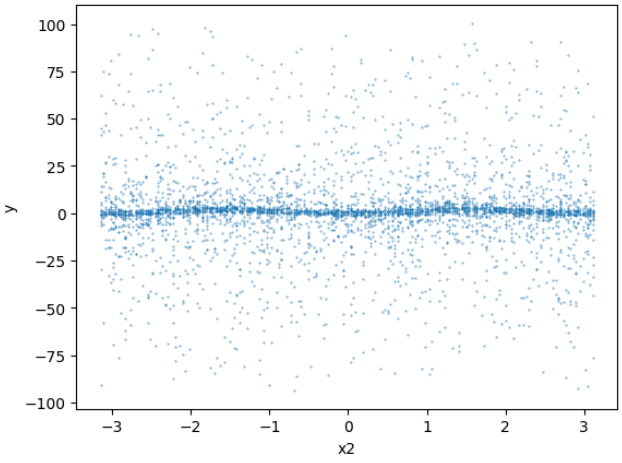
We apply PAWN to one of the most frequently used benchmark models in the Sensitivity Analysis literature, the Ishigami-Homma function:

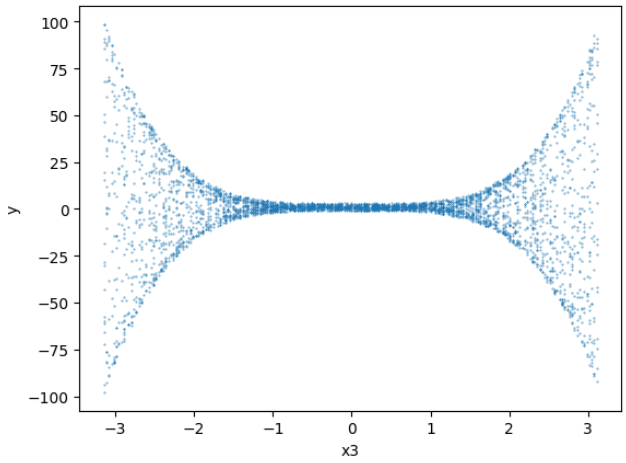
$$y = \sin(x_1) + a\,\sin(x_2)^2 + bx_3^4\,\sin(x_1)$$

where all xi follow a uniform distribution over (-pi,+pi)and a = 2 and b = 1. The figures below report the scatter plots of the output against the three inputs. It can be noticed that:

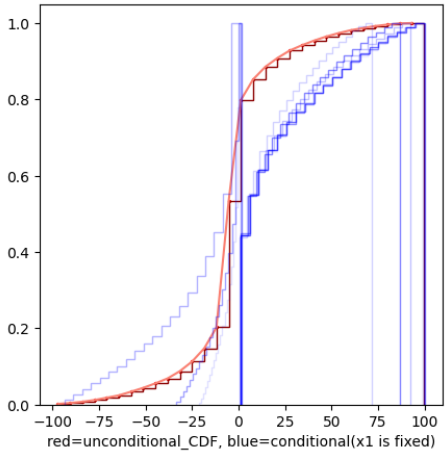
(i) x1 seems to be the most influential input;
(ii) x2 seems to be non-influential.

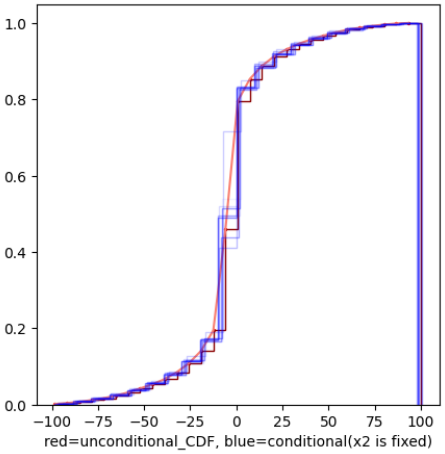


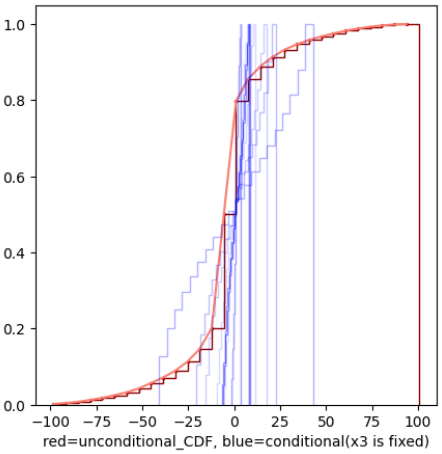




If we apply our PAWN approach we obtain the three sets of conditional CDFs. In all panels, the red line is the empirical unconditional output CDF, while the blue lines are the conditional CDFs obtained at n=10 different conditioning values of xi.

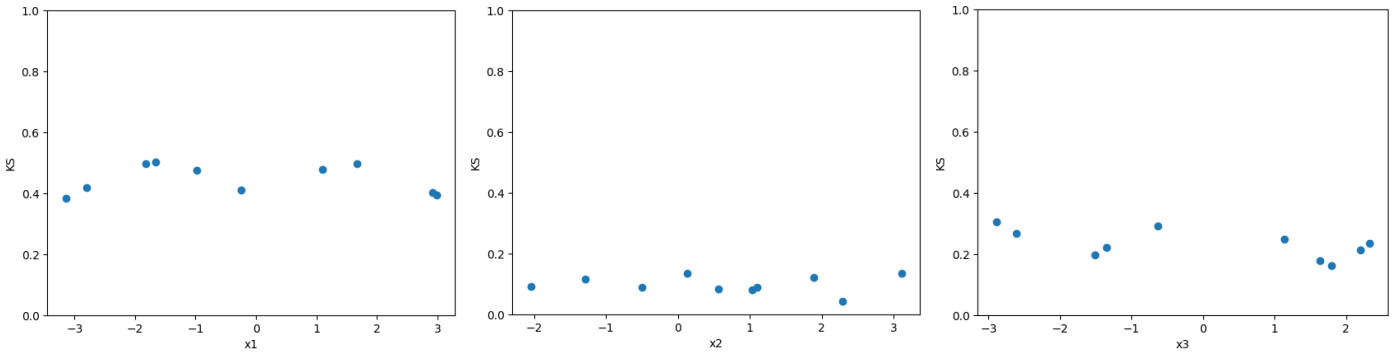






Visual analysis of the CDFs immediately shows that x1 and x3 are both much more influential than x2, as their conditional CDFs are more widespread around the unconditional one (red line), while those of x2 almost overlap with the unconditional CDF.

By computing Kolmogorov-Smirnov statistic between unconditional and conditional CDFs below plots are obtained:



If we consider median as statistic the PAWN sensitivity indices are equal to:

```
PAWN Sensivity indexes:
median_kstest_values1= 0.44785
median_kstest_values2= 0.09003333333333335
median_kstest_values3= 0.22751666666666664
```

The implementation python code is available by this link: Sensitivity_analysis_PAWN.ipynb

Numerical example 2:

Also we apply PAWN to the abc model for modeling relationships among precipitation, evapo-transpiration, groundwater storage, and streamflow using only three model parameters:

$$Q_t = (1 - a - b)P_t + cG_{t-1} + \varepsilon_t$$

$$G_t = (1 - c)G_{t-1} + aP_t + \nu_t$$

Since the parameters represent fractions they have upper and lower limits 0<a,b,c<1, and since infiltration and evapotranspiration combined cannot exceed total precipitation, 0<a+b<1.

A,b,c has uniform distribution. We fit different distribution on precipitation data and find that betta distribution is best fit.

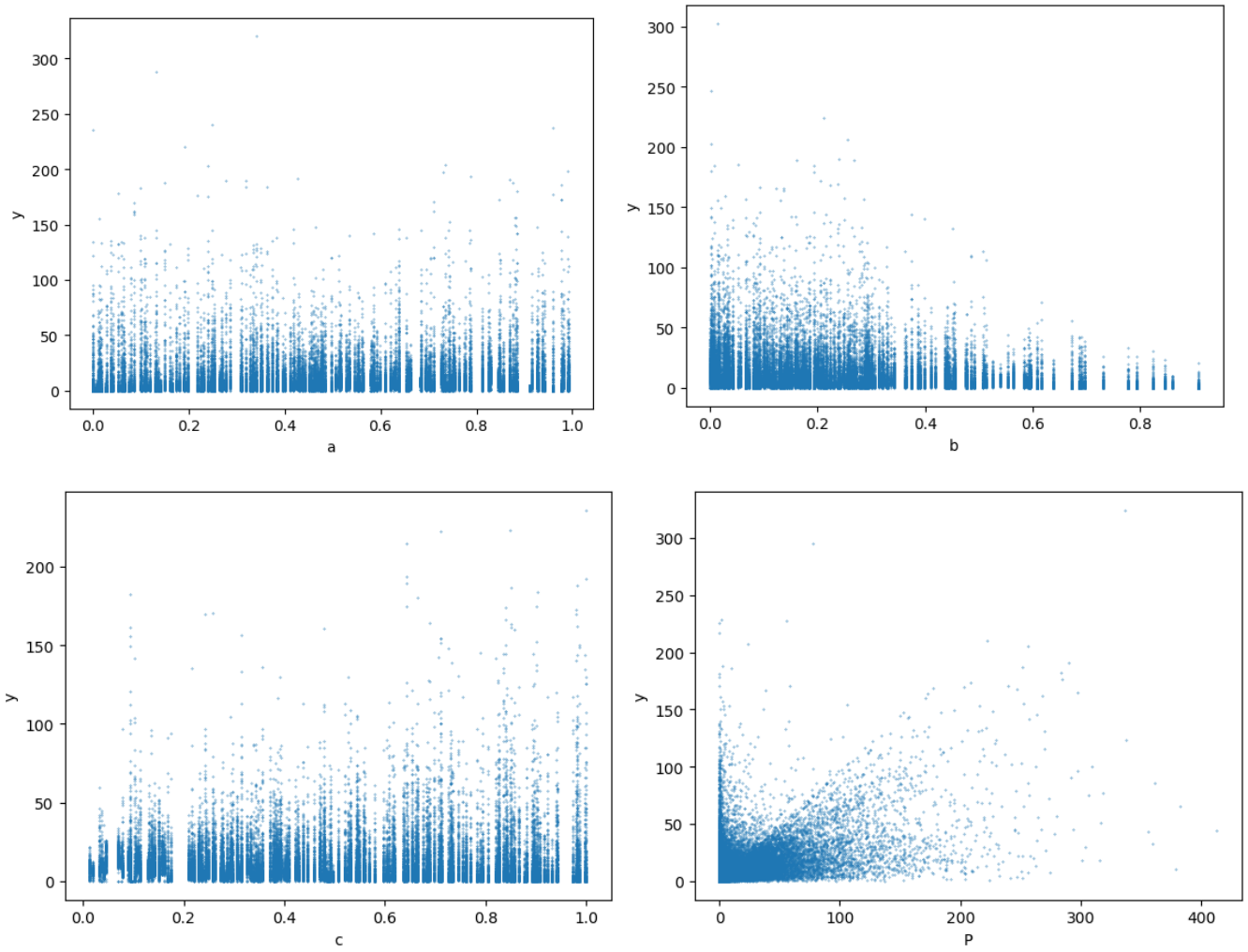
	sumsquare_error	aic	bic	kl_div	ks_statistic	ks_pvalue
beta	0.002420	1890.201179	-4264.057811	inf	0.350201	5.655773e-40
chi2	0.003270	11267.916444	-4161.558652	inf	0.543708	2.938654e-100
exponpow	0.003326	1472.282525	-4155.511864	inf	0.336111	8.988220e-37
lognorm	0.005312	1613.544526	-3986.960982	inf	0.336111	8.987982e-37
gamma	0.005535	1646.279730	-3972.130339	inf	0.336111	8.987982e-37

```
{'beta': {'a': 0.23988695969936633,
  'b': 24.053310038119534,
  'loc': -2.76126312081761e-30,
  'scale': 1543.0547732112127}}
```

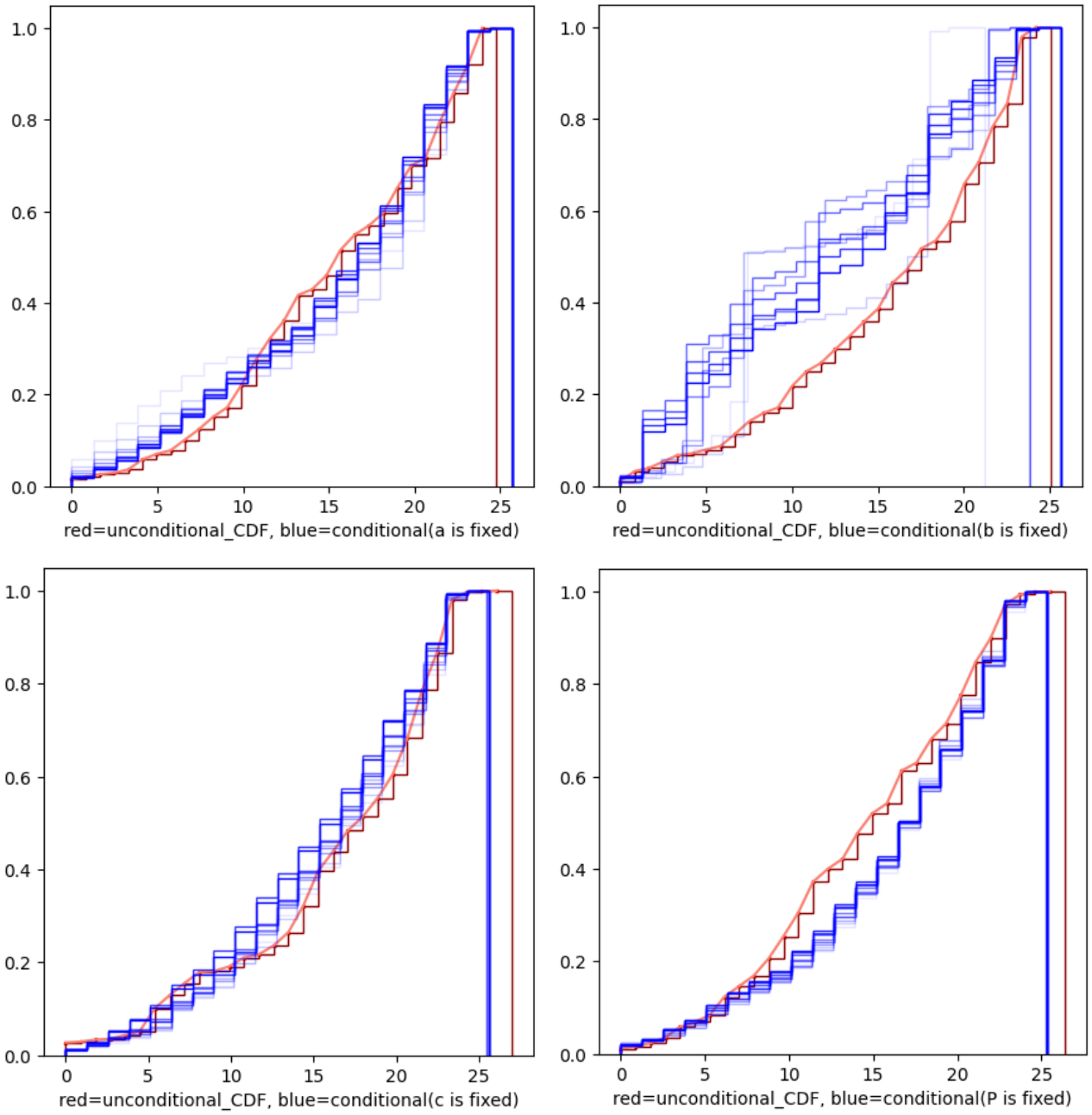
The figures below reports the scatter plots of the output against the P,a,b,c inputs that is produced randomly based on their distributions . It can be noticed that:

(i) b and p seem to be the most influential inputs;

(ii) a and c seem to be less influential.

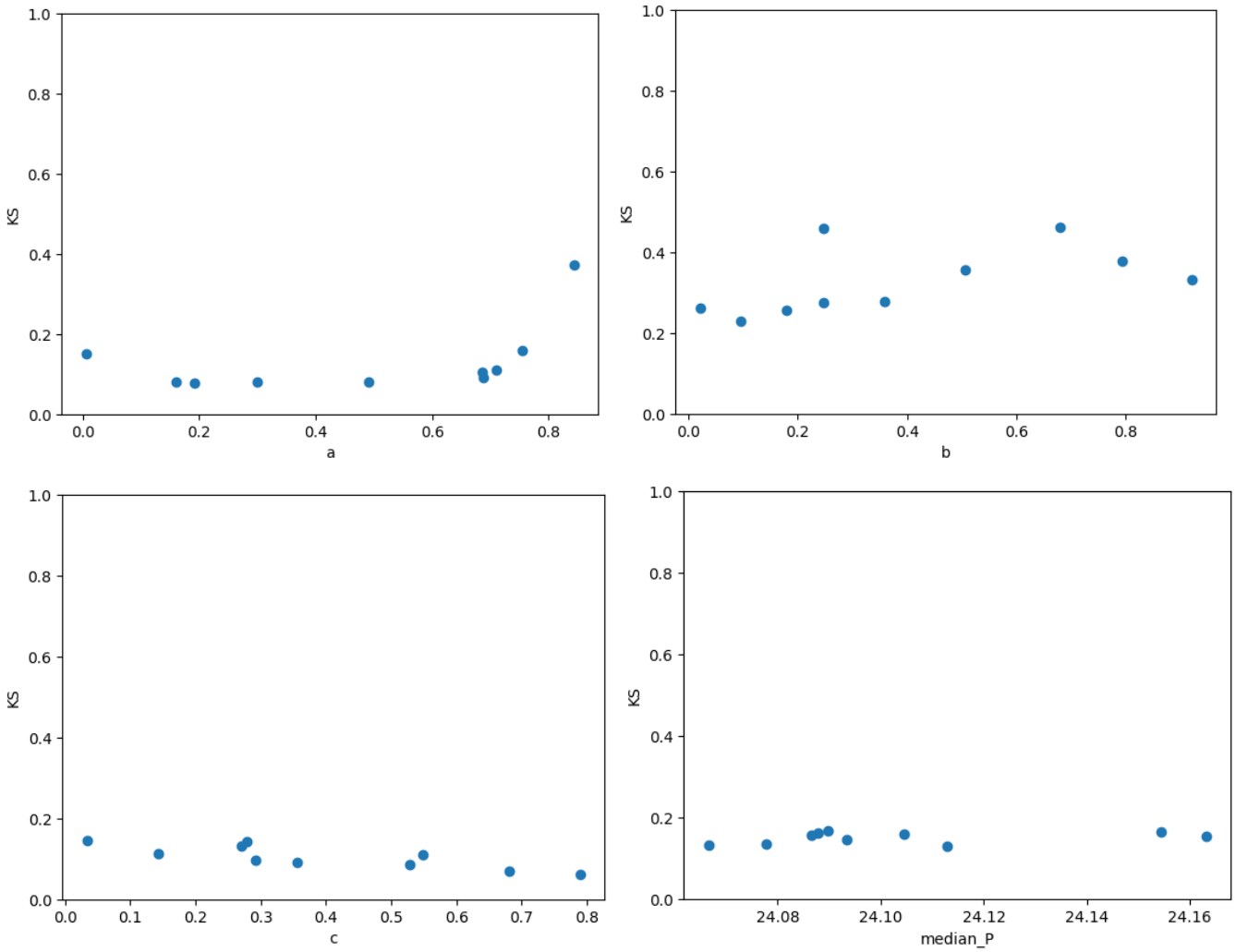


If we apply our PAWN approach, we obtain the four sets of conditional CDFs. In all panels, the red line is the empirical unconditional output CDF, while the blue lines are the conditional CDFs obtained at n=10 different conditioning values of a, b, c and P.



Visual analysis of the CDFs shows b is most influential and then P is more influential than two other inputs, as their conditional CDFs are more widespread around the unconditional one (red line), while those of a and c almost overlap with the unconditional CDF.

By computing Kolmogorov-Smirnov statistic between unconditional and conditional CDFs below plots are obtained:



If we consider median as statistic the PAWN sensitivity indices are equal to:

```
PAWN Sensivity indexes
median_kstest_values_a= 0.09881909142433741
median_kstest_values_b= 0.30405498100415773
median_kstest_values_c= 0.10294501772911697
median_kstest_values_P= 0.15453372509507346
```

The implementation python code is available by this link:[pawn_for_abc.ipynb]

Appendix:

Kolmogorov–Smirnov statistic

The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples.

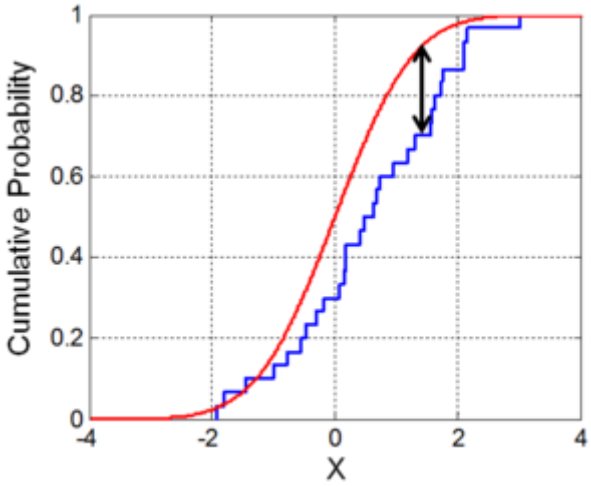


Figure 4: Illustration of the Kolmogorov–Smirnov statistic. The red line is a model CDF, the blue line is an empirical CDF, and the black arrow is the KS statistic

Kolmogorov–Smirnov test

In statistics, the Kolmogorov–Smirnov test (K–S test or KS test) is a nonparametric test of the equality of continuous (or discontinuous)one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test).

References:

- Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., & Wagener, T. (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. Environmental Modelling & Software, 79, 214-232. https://doi.org/https://doi.org/10.1016/j.envsoft.2016.02.008
- Pianosi, F., & Wagener, T. (2015). A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. Environmental Modelling & Software, 67, 1-11. https://doi.org/https://doi.org/10.1016/j.envsoft.2015.01.004
- Reed, Patrick & Hadjimichael, Antonia & Malek, Keyvan & Karimi, Tina & vernon, chris & Srikrishnan, Vivek & Gupta, Rohini & Gold, David & Lee, Ben & Keller, Klaus & thurber, travis. (2022). Addressing Uncertainty in MultiSector Dynamics Research. 10.5281/zenodo.6110623.