Surrogating the response PDF of stochastic simulators using parametric & semi-parametric representations

Conference Poster

Author(s):
Zhu, Xujia; Sudret, Bruno

Publication date:
2018-03-21

Permanent link:
https://doi.org/10.3929/ethz-b-000309591

![ETH zürich]

Surrogating the Response PDF of Stochastic Simulators Using Parametric & Semi-parametric Representations

X. Zhu, B. Sudret

ETH Zürich, Chair of Risk, Safety and Uncertainty Quantification, zhu@ibk.baug.ethz.ch

Stochastic simulators & emulators

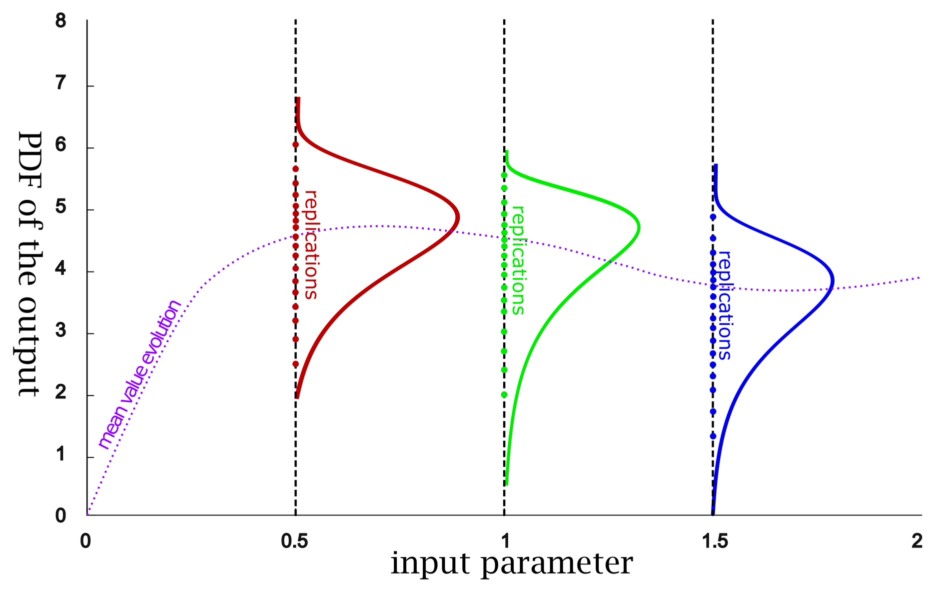
Stochastic simulators

- *Stochastic simulators* provide different results when run with the same input several times. Therefore, they are in nature random processes with the input parameters as index (assume second order process).

$$\mathcal{M} : \mathcal{D}_{\boldsymbol{x}} \subset \mathbb{R}^M \to L^2(\Omega, \mathcal{F}, \mathbb{P}) : \boldsymbol{x} \mapsto Y(\boldsymbol{x})$$

Source of the randomness

- The stochastic simulator can be considered as a deterministic mapping $(\boldsymbol{x}, \boldsymbol{z}) \mapsto \mathcal{M}_{det}(\boldsymbol{x}, \boldsymbol{z})$, where $\boldsymbol{z} \in \mathcal{D}_{\boldsymbol{z}}$ are hidden variables.



Goal of the stochastic emulator

- Based on the data at design points, predict the PDF with a new set of input parameters.

General methodology

- Estimation of the PDF at $\boldsymbol{x}_i$ from *replications*.
- Parametrization of the PDF.
- Prediction of the PDF for a new $\boldsymbol{x}$ using Gaussian process modelling.

Marrel, A. et al. (2012). *Global sensitivity analysis of stochastic computer models with joint metamodels.* Stat. Comput. 22, 833-847.

Moutoussamy, V. et al. (2015). *Emulators for stochastic simulation codes.* ESAIM: Mathematical Modelling and Numerical Analysis 48, 116-155.

Parametric approach

Assumptions

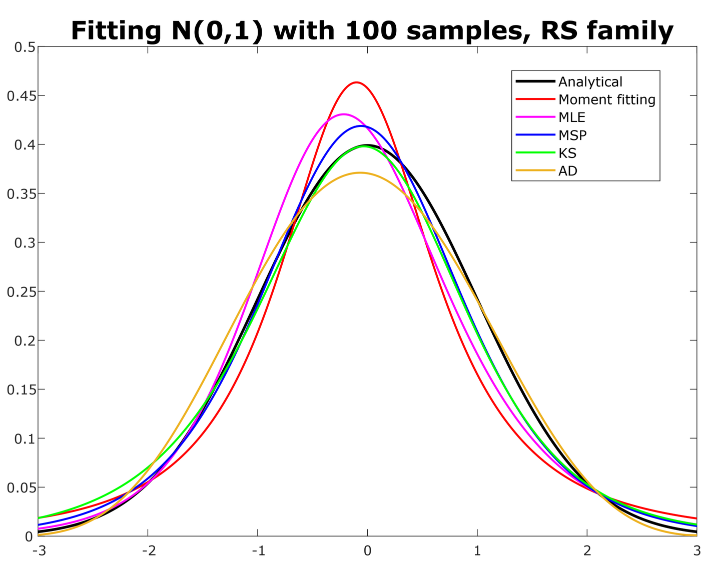
- The model response distributions belong to a parametric family.
- Distribution parameters vary smoothly with respect to $\boldsymbol{x}$.

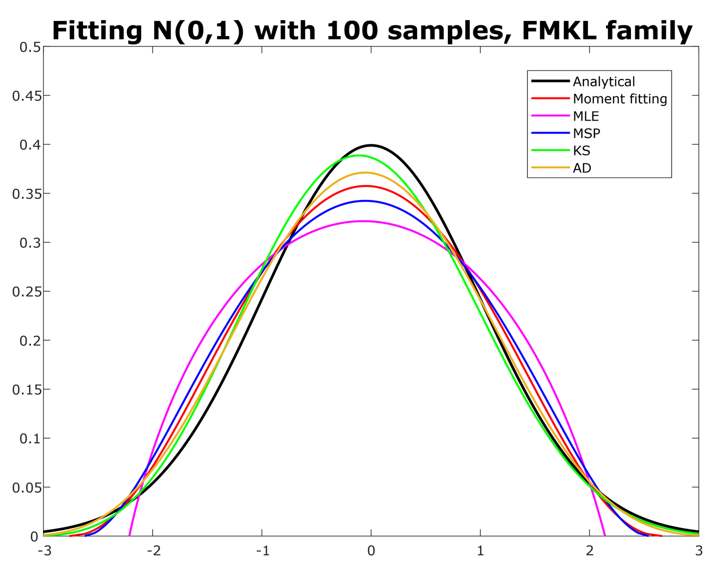
Generalized Lambda distribution

- The *GLD* can approximate many well-known distribution families such as normal, exponential, lognormal, Weibull, Student's t, beta, gamma, chi-square, etc.
- The *quantile function* is parametrized. There are two widely used families:
  - The Ramberg and Schmeiser (*RS*) family
  $$Q(u) = \lambda_1 + \frac{1}{\lambda_2}\left(u^{\lambda_3} - (1-u)^{\lambda_4}\right)$$
  - The Freimer, Mudholkar, Kollia, and Lin (*FMKL*) family
  $$Q(u) = \lambda_1 + \frac{1}{\lambda_2}\left(\frac{u^{\lambda_3}-1}{\lambda_3} - \frac{(1-u)^{\lambda_4}-1}{\lambda_4}\right)$$
- The *RS* family has a complicated feasible domain for $\boldsymbol{\lambda}$ (6 zones), while that of the *FMKL* family is simple (only one single zone).

Estimation methods

- Moment fitting
- Maximum likelihood estimation (*MLE*) or spacing estimation (*MSP*)
- Minimization of the Kolmogorov-Smirnov (*KS*) or the Anderson–Darling (*AD*) statistics.





Gaussian process modelling

Gaussian process modelling (a.k.a Kriging) is an interpolation algorithm assuming that the data is a realization of a Gaussian process.

$$\mathcal{M}^K(x) = \boldsymbol{\beta}^T \cdot \boldsymbol{f}(x) + \sigma^2 \cdot Z(x, \omega)$$

- Choose the trend functions $\boldsymbol{f}(x)$ and the type of covariance function corresponding to the Gaussian process $Z(x, \omega)$.
- Estimate the hyperparameters $\boldsymbol{\theta}$ associated to the covariance function via MLE/RMLE(restricted maximum likelihood) or cross-validation (leave-one-out error).
- Compute the Kriging parameters $\boldsymbol{\beta}$ and σ^2 resorting to MLE/RMLE.
- Calculate the conditional probability with a new set of parameters and use the mean value as the predicted value.

Toy example

Description of the example

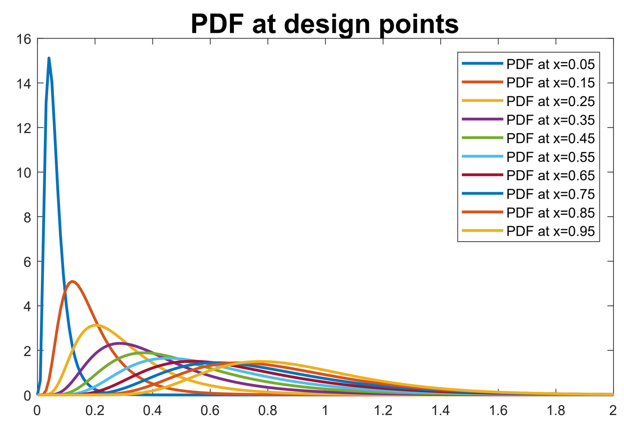
- $Y(x) = \mathcal{M}_{det}(x, \boldsymbol{Z}) = \sin(x) \cdot (Z_1 \cdot Z_2)^{\cos(x)}$
- $\mathcal{D}_x = [0, 1]$ and Z_1, Z_2 are independent lognormal random variables
- The output with each given x follows also a lognormal distribution

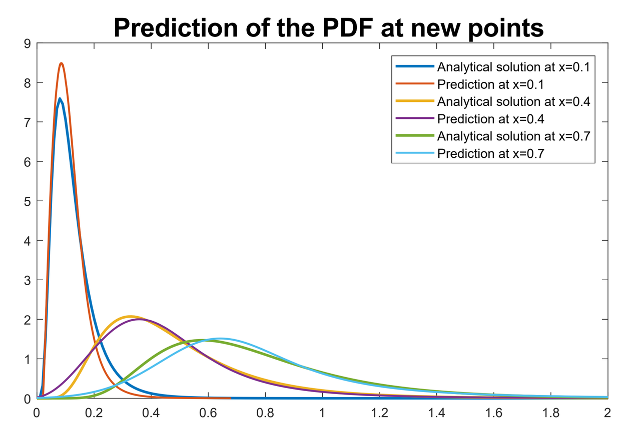
Data generation

- At each design point (10 points on a regular grid from 0.05 to 0.95), 100 independent runs of the function.

Estimation details

- Estimate the PDF at each design point using *GLD*. Note that the initial values of $\boldsymbol{\lambda}$ at x_i are set to be the results $\boldsymbol{\lambda}(x_{i-1})$.
- Predict *independently* the 4 parameters $\boldsymbol{\lambda}(x)$ applying Kriging.





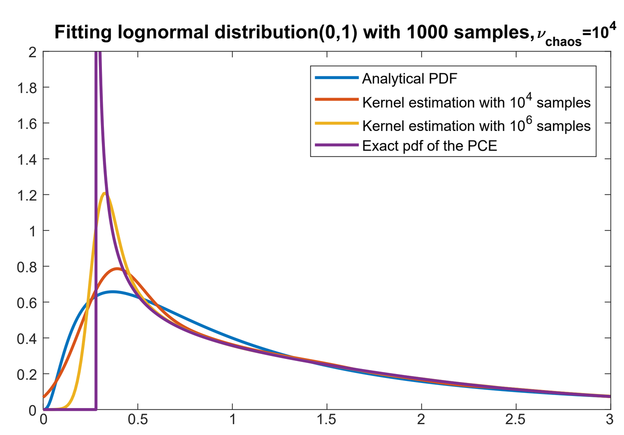
Average relative error of quantiles

Quantile	Q5	Q10	Q50	Q90	Q95
10 design points	11.59%	8.39%	3.38%	7.28%	10.66%
80 new points	9.19%	7.14%	3.34%	8.14%	11.84%

The PCE-based semi-parametric PDF estimation

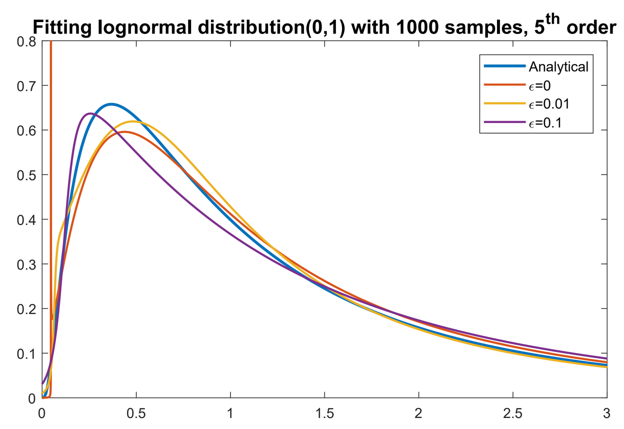
Overview of the method

- $Y \stackrel{d}{=} g(\Xi) = \sum_{i=1}^{N_P} a_i h_i(\Xi)$ where Ξ follows the standard normal distribution and $h_i(\Xi)$ are Hermite polynomials.
- Potential problems:
  - There exists infinite number of $(a_i)_{i=1,\ldots,N_P}$ that can lead the likelihood to infinity.
  - Different $(a_i)_{i=1,\ldots,N_P}$ can result in exactly the same distribution.



Improvement and discussion

- Restriction to strictly increasing polynomials g with $g'(x) > \epsilon$ over $(-\infty, +\infty)$.
- According to some tests, the estimated PDF is very sensitive to the choice of ϵ. Its optimal value depends strongly on the data, especially when the amount of data is small.



Conclusion and outlook

- The PDF estimation error is dominant in most cases because of the limited number of *replications*.
- Generalized lambda distribution gives quite convincing results according to the toy example.
- More robust techniques are needed to improve the PCE-based semi-parametric PDF estimation.

Chair of Risk, Safety and Uncertainty Quantification
Stefano-Franscini-Platz 5
CH-8093 Zürich

Risk, Safety & Uncertainty Quantification