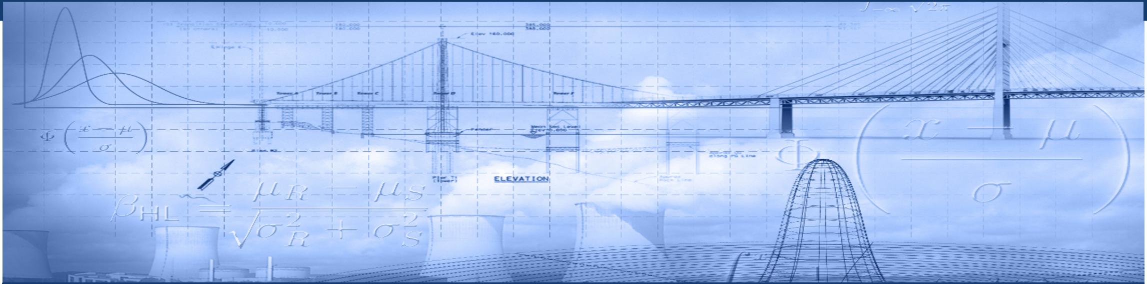
ETH zürich



Emulating the response distribution of stochastic simulators

Xujia Zhu

Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich

April 28, 2021 – MASCOTNUM 2021 Meeting

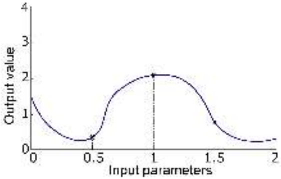
Outline

Deterministic vs. stochastic simulators

Deterministic simulators

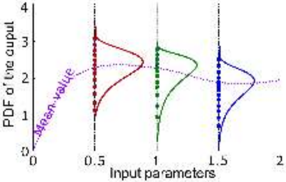
- Each set of input variables has a unique corresponding output

$$\mathcal{M}_d : \mathcal{D}_{\boldsymbol{X}} \subset \mathbb{R}^M \to \mathbb{R}$$



Stochastic simulators

- A given set of input parameters can lead to different values of the output
- $Y(\boldsymbol{x})$ is a random variable
- Source of randomness: $Y(\boldsymbol{x}) = \mathcal{M}(\boldsymbol{x}, \boldsymbol{Z})$, where $\boldsymbol{Z}$ are latent variables



Computational costs induced by stochastic simulators

- Replications are needed to estimate the PDF of $Y(\boldsymbol{x})$ (i.e., $Y \mid \boldsymbol{X} = \boldsymbol{x}$)
- Many runs must be carried out by varying $\boldsymbol{X}$ for uncertainty propagation, sensitivity analysis, optimization, etc.
- Realistic simulators (e.g., for wind turbine design) are costly

Need for surrogate models

- Non-intrusive (i.e., that considers the stochastic simulator as a black box)
- General-purpose: no restrictive assumption (e.g., Gaussian) on the family of the output
- Able to tackle the full distribution of $Y(\boldsymbol{x})$, but also quantities of interest (e.g., mean, variance, quantiles)
- Providing a representation of $Y(\boldsymbol{x})$ easy to sample from

Computational costs induced by stochastic simulators

- Replications are needed to estimate the PDF of $Y(\boldsymbol{x})$ (i.e., $Y \mid \boldsymbol{X} = \boldsymbol{x}$)
- Many runs must be carried out by varying $\boldsymbol{X}$ for uncertainty propagation, sensitivity analysis, optimization, etc.
- Realistic simulators (e.g., for wind turbine design) are costly

Need for surrogate models

- Non-intrusive (i.e., that considers the stochastic simulator as a black box)
- General-purpose: no restrictive assumption (e.g., Gaussian) on the family of the output
- Able to tackle the full distribution of $Y(\boldsymbol{x})$, but also quantities of interest (e.g., mean, variance, quantiles)
- Providing a representation of $Y(\boldsymbol{x})$ easy to sample from

Computational costs induced by stochastic simulators

- Replications are needed to estimate the PDF of $Y(\boldsymbol{x})$ (i.e., $Y \mid \boldsymbol{X} = \boldsymbol{x}$)
- Many runs must be carried out by varying $\boldsymbol{X}$ for uncertainty propagation, sensitivity analysis, optimization, etc.
- Realistic simulators (e.g., for wind turbine design) are costly

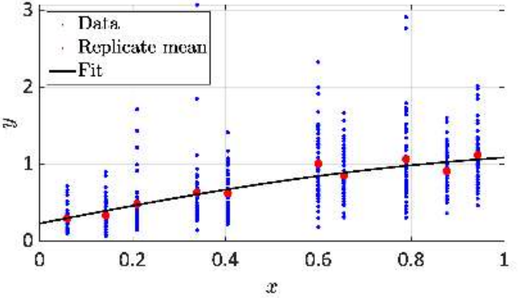
Need for surrogate models

- Non-intrusive (i.e., that considers the stochastic simulator as a black box)
- General-purpose: no restrictive assumption (e.g., Gaussian) on the family of the output
- Able to tackle the full distribution of $Y(\boldsymbol{x})$, but also quantities of interest (e.g., mean, variance, quantiles)
- Providing a representation of $Y(\boldsymbol{x})$ easy to sample from

Outline

Existing methods

- Replication-based:
  - Quantile estimation: Plumlee & Tuo (2014) *Building accurate emulators for stochastic simulations via quantile Kriging*, Technometrics
  - Kernel smoothing: Moutoussamy *et al.* (2015) *Emulators for stochastic simulation codes*, ESAIM: Math. Model. Num. Anal.



Existing methods

- Replication-based:
  - Quantile estimation: Plumlee & Tuo (2014) *Building accurate emulators for stochastic simulations via quantile Kriging*, Technometrics
  - Kernel smoothing: Moutoussamy *et al.* (2015) *Emulators for stochastic simulation codes*, ESAIM: Math. Model. Num. Anal.
- Random field representation $Y_{\boldsymbol{x}}(\omega) = \mathcal{M}\left(\boldsymbol{x}, \boldsymbol{Z}(\omega)\right)$: Azzi *et al.* (2019) *Surrogate modeling of stochastic functions - application to computational electromagnetic dosimetry*, Int. J. Uncertainty Quantification

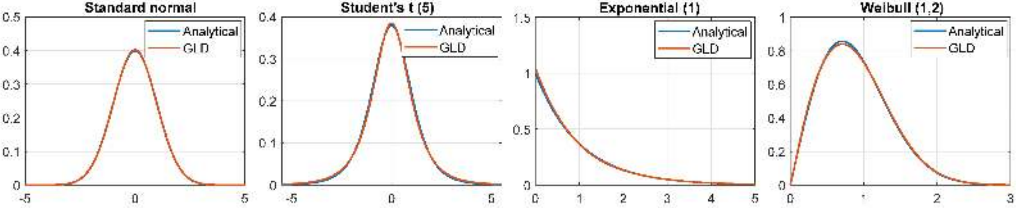
Existing methods

- Replication-based:
  - Quantile estimation: Plumlee & Tuo (2014) *Building accurate emulators for stochastic simulations via quantile Kriging*, Technometrics
  - Kernel smoothing: Moutoussamy *et al.* (2015) *Emulators for stochastic simulation codes*, ESAIM: Math. Model. Num. Anal.
- Random field representation $Y_{\boldsymbol{x}}(\omega) = \mathcal{M}\left(\boldsymbol{x}, \boldsymbol{Z}(\omega)\right)$: Azzi *et al.* (2019) *Surrogate modeling of stochastic functions - application to computational electromagnetic dosimetry*, Int. J. Uncertainty Quantification
- Statistical approach:
  - Under the assumption of normality:
    Marrel *et al.* (2012) *Global sensitivity analysis of stochastic computer models with joint metamodels*, Stat. Comput.
    Binois *et al.* (2018) *Practical heteroscedastic Gaussian process modeling for large simulation experiments*, J. Comput. Graph. Stat.
  - Quantile regression: Koenker & Bassett (1978) *Regression quantiles*, Econometrica: journal of the Econometric Society
  - Kernel smoothing: Hall *et al.* (2004) *Cross-validation and the estimation of conditional probability densities*, J. Amer. Stat. Assoc.

Outline

Generalized lambda distributions

- Flexibility: able to approximate most of the parametric distributions



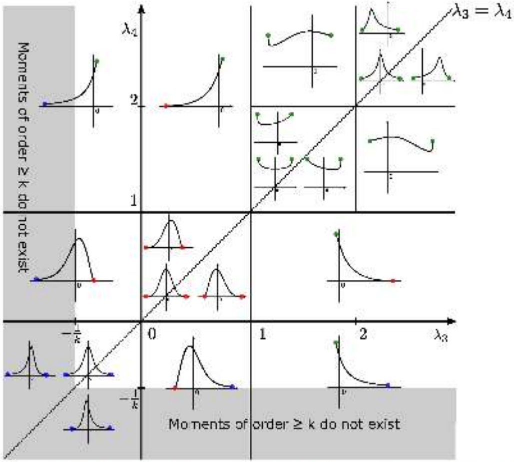
- The Freimer-Mudholkar-Kollia-Lin (FMKL) lambda distribution is defined through its quantile function

$$Q(u;\boldsymbol{\lambda}) = \lambda_1 + \frac{1}{\lambda_2}\left(\frac{u^{\lambda_3}-1}{\lambda_3} - \frac{(1-u)^{\lambda_4}-1}{\lambda_4}\right)$$

- The PDF is obtained by:

$$f_Y(y;\boldsymbol{\lambda}) = \frac{1}{Q'(u;\boldsymbol{\lambda})} = \frac{\lambda_2}{u^{\lambda_3-1}+(1-u)^{\lambda_4-1}} \qquad \text{with } u = Q^{-1}(y;\boldsymbol{\lambda})$$

Properties



- λ_3 and λ_4 control the shape and boundedness

$$B_l(\boldsymbol{\lambda}) = \begin{cases} -\infty, & \lambda_3 \leq 0 \\ \lambda_1 - \frac{1}{\lambda_2 \lambda_3}, & \lambda_3 > 0 \end{cases}$$

$$B_u(\boldsymbol{\lambda}) = \begin{cases} +\infty, & \lambda_4 \leq 0 \\ \lambda_1 + \frac{1}{\lambda_2 \lambda_4}, & \lambda_4 > 0 \end{cases}$$

- Blue points: infinite support
- Red points: finite support, with PDF = 0 at the bound
- Green points: finite support, with PDF $\neq 0$ at the bound

Zhu & Sudret (2020). *Replication-based emulation of the response distribution of stochastic simulators using generalized lambda distributions*, Int. J. Uncertainty Quantification, 10:249–275

Generalized lambda models (GLaM)

General setting

$$Y(\boldsymbol{x}) \sim \mathrm{GLD}\left(\lambda_1(\boldsymbol{x}), \lambda_2(\boldsymbol{x}), \lambda_3(\boldsymbol{x}), \lambda_4(\boldsymbol{x})\right)$$

Polynomial chaos expansions

$$\lambda_k(\boldsymbol{x}) = \lambda_k^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d} c_{k,\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) \quad k = 1, 3, 4$$

$$\lambda_2(\boldsymbol{x}) = \lambda_2^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c}) = \exp\left(\sum_{\boldsymbol{\alpha} \in \mathbb{N}^d} c_{2,\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{x})\right)$$

- Independent input parameters with $\boldsymbol{X} \sim f_{\boldsymbol{X}} = \prod_{j=1}^{d} f_{X_j}$
- Basis functions (multivariate polynomials) $\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \prod_{j=1}^{d} \phi_{\alpha_j}^{(j)}(x_j)$
- $\boldsymbol{c}$ are the model parameters to be estimated

Generalized lambda models (GLaM)

General setting

$$Y(\boldsymbol{x}) \sim \mathrm{GLD}\left(\lambda_1(\boldsymbol{x}), \lambda_2(\boldsymbol{x}), \lambda_3(\boldsymbol{x}), \lambda_4(\boldsymbol{x})\right)$$

Polynomial chaos expansions

$$\lambda_k(\boldsymbol{x}) \approx \lambda_k^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}_k} c_{k,\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) \quad k = 1, 3, 4$$

$$\lambda_2(\boldsymbol{x}) \approx \lambda_2^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c}) = \exp\left(\sum_{\boldsymbol{\alpha} \in \mathcal{A}_2} c_{2,\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{x})\right)$$

- Independent input parameters with $\boldsymbol{X} \sim f_{\boldsymbol{X}} = \prod_{j=1}^{d} f_{X_j}$
- Basis functions (multivariate polynomials) $\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \prod_{j=1}^{d} \phi_{\alpha_j}^{(j)}(x_j)$
- $\boldsymbol{c}$ are the model parameters to be estimated

Estimation with given PCE basis

Data generation

- Experimental design of size N in the $\boldsymbol{X}$-space: $\mathcal{X} = \left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\right\}$
- The simulator is evaluated *once*, i.e., no replications needed, for each $\boldsymbol{x}^{(i)} \in \mathcal{X}$: $y^{(i)} \overset{\text{def}}{=} \mathcal{M}\left(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}\right)$

Idea

- Build a global model for the joint distribution of inputs and outputs:

$$f_{\boldsymbol{X},Y}(\boldsymbol{x}, y) = f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) \cdot f_{\boldsymbol{X}}(\boldsymbol{x})$$

  where the conditional PDF is represented by a generalized lambda model:

$$f^{\mathrm{GLD}}_{\boldsymbol{X},Y}(\boldsymbol{x}, y;\, \boldsymbol{c}) = f^{\mathrm{GLD}}_{Y|\boldsymbol{X}}\left(y; \boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c})\right) \cdot f_{\boldsymbol{X}}(\boldsymbol{x})$$

- Find the optimal PCE coefficients $\boldsymbol{c}^*$ that minimize the Kullback-Leibler divergence between $f_{\boldsymbol{X},Y}(\boldsymbol{x}, y)$ and $f^{\mathrm{GLD}}_{\boldsymbol{X},Y}(\boldsymbol{x}, y)$:

$$\boldsymbol{c}^* = \arg\min_{\boldsymbol{c}} D_{\mathrm{KL}}\left(f_{\boldsymbol{X},Y} \,||\, f^{\mathrm{GLD}}_{\boldsymbol{X},Y}(\,\cdot\,; \boldsymbol{c})\right)$$

Estimation with given PCE basis (cont.)

Maximum likelihood estimation

- The minimization problem is equivalent to

$$\boldsymbol{c}^* = \arg\max_{\boldsymbol{c}} \mathbb{E}_{\boldsymbol{X},Y}\left[\log f^{\mathrm{GLD}}_{Y|\boldsymbol{X}}\left(Y; \boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{X};\boldsymbol{c})\right)\right]$$

- Maximum likelihood estimator

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} \frac{1}{N}\sum_{i=1}^{N} \log f^{\mathrm{GLD}}_{Y|\boldsymbol{X}}\left(y^{(i)}; \boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x}^{(i)};\boldsymbol{c})\right)$$

- Consistency: if the simulator is a GLaM for $\boldsymbol{c}^*$, under mild conditions $\hat{\boldsymbol{c}} \xrightarrow{\text{a.s.}} \boldsymbol{c}^*$ as $N \to +\infty$

Zhu & Sudret (2021) *Emulation of stochastic simulators using generalized lambda models*, Submitted to SIAM/ASA J. Unc. Quant.

Estimation with unknown PCE basis

With replications

- R replications for each $\boldsymbol{x}^{(i)} \in \mathcal{X}$: $\mathcal{Y}^{(i)} = \left\{y^{(i,1)}, y^{(i,2)}, \dots, y^{(i,R)}\right\}$
- Infer a generalized lambda distribution $\hat{\boldsymbol{\lambda}}^{(i)}$ for each point $\boldsymbol{x}^{(i)}$ of the experimental design based on the replications $\mathcal{Y}^{(i)}$
- Fit a sparse polynomial chaos expansion to the parameters $\left\{\left(\boldsymbol{x}^{(1)}, \hat{\boldsymbol{\lambda}}^{(1)}\right), \dots, \left(\boldsymbol{x}^{(N)}, \hat{\boldsymbol{\lambda}}^{(N)}\right)\right\}$, which selects the basis functions for $\boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x})$
- MLE with all the data to estimate the coefficients

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} \frac{1}{NR} \sum_{i=1}^{N} \sum_{r=1}^{R} \log f^{\mathrm{GLD}}_{Y|\boldsymbol{X}}\left(y^{(i,r)};\, \boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x}^{(i)}; \boldsymbol{c})\right)$$

Zhu & Sudret (2020), *Replication-based emulation of the response distribution of stochastic simulators using generalized lambda distributions*, Int. J. Uncertainty Quantification, 10:249–275

Estimation with unknown PCE basis

Without replications

- PCE models for the mean and variance of the model output built using the feasible generalized least-square method
- Use the PCE basis of $\mu(\boldsymbol{x})$ (resp. $\log \sigma^2(\boldsymbol{x})$) for λ_1 (resp. λ_2)
- PCE of degree 1 for λ_3 and λ_4 (it is assumed that the shape of the response distribution does not vary nonlinearly with $\boldsymbol{x}$)
- MLE to estimate the coefficients

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} \frac{1}{N} \sum_{i=1}^{N} \log f^{\text{GLD}}_{Y|\boldsymbol{X}} \left(y^{(i)}; \boldsymbol{\lambda}^{\text{PC}}(\boldsymbol{x}^{(i)}; \boldsymbol{c}) \right)$$

Zhu & Sudret (2021) *Emulation of stochastic simulators using generalized lambda models*, Submitted to SIAM/ASA J. Unc. Quant.

Outline

Motivation

Another perspective of GLaM

$$Y(\boldsymbol{x}) \overset{\mathrm{d}}{=} F^{-1}_{Y|\boldsymbol{X}}(U \mid \boldsymbol{x}) \approx Q^{\mathrm{GLD}}\left(U; \boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c})\right)$$

- The variable $U \sim \mathcal{U}(0,1)$ can be seen as the source of stochasticity, and the quantile transform represents the model response
- This is a stochastic surrogate: when fixing $\boldsymbol{x}$ and sampling U, one obtains samples for the surrogate model response

Latent variable model

- Represent the model response as a transform of a latent variable $\tilde{Z}$, e.g., $Y(x) \overset{\mathrm{d}}{\approx} g(\tilde{Z}; x)$
- Stochastic PCE: the transform is given by a PCE

Motivation

Another perspective of GLaM

$$Y(\boldsymbol{x}) \stackrel{\mathrm{d}}{=} F^{-1}_{Y|\boldsymbol{X}}(U \mid \boldsymbol{x}) \approx Q^{\mathrm{GLD}}\left(U; \boldsymbol{\lambda}^{\mathrm{PC}}(\boldsymbol{x}; \boldsymbol{c})\right)$$

- The variable $U \sim \mathcal{U}(0,1)$ can be seen as the source of stochasticity, and the quantile transform represents the model response
- This is a stochastic surrogate: when fixing $\boldsymbol{x}$ and sampling U, one obtains samples for the surrogate model response

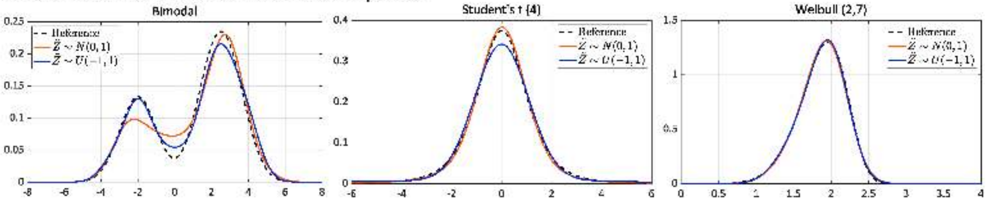
Latent variable model

- Represent the model response as a transform of a latent variable $\tilde{Z}$, e.g., $Y(\boldsymbol{x}) \stackrel{\mathrm{d}}{\approx} g(\tilde{Z}; \boldsymbol{x})$
- Stochastic PCE: the transform is given by a PCE

Formulation

$$Y(\boldsymbol{x}) \overset{\mathrm{d}}{\approx} \sum_{\alpha \in \mathcal{A}} c_\alpha \Psi_\alpha(\boldsymbol{x}, \tilde{Z}) + \epsilon$$

- $\tilde{Z}$ is a latent variable, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a noise variable
- $\tilde{Z}$ and ϵ are introduced to represent the random nature of the stochastic simulator: for a given $\boldsymbol{x}$, $Y(\boldsymbol{x})$ is a function of the latent variable $\tilde{Z}$ plus ϵ



Formulation

$$Y(\boldsymbol{x}) \overset{\mathrm{d}}{\approx} \sum_{\boldsymbol{\alpha}\in\mathcal{A}} c_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}, \tilde{Z}) + \epsilon$$

- $\tilde{Z}$ is a latent variable, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a noise variable
- $\tilde{Z}$ and ϵ are introduced to represent the random nature of the stochastic simulator: for a given $\boldsymbol{x}$, $Y(\boldsymbol{x})$ is a function of the latent variable $\tilde{Z}$ plus ϵ
- By convolution, the response distribution is given by

$$f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) = \int_{\mathcal{D}_{\tilde{Z}}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y - \sum_{\boldsymbol{\alpha}\in\mathcal{A}} c_{\boldsymbol{\alpha}} \Psi(\boldsymbol{x}, \tilde{z})\right)^2}{2\sigma^2}\right) f_{\tilde{Z}}(\tilde{z}) \mathrm{d}\tilde{z}$$

- To build a stochastic PCE, $\boldsymbol{c}$ and σ should be estimated from data

Estimation method

Maximum likelihood estimation

- The conditional likelihood for a data point $(\boldsymbol{x}, y)$ is

$$l(\boldsymbol{c}, \sigma; \boldsymbol{x}, y) = \int_{\mathcal{D}_{\tilde{Z}}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y - \sum_{\boldsymbol{\alpha}\in\mathcal{A}} c_{\boldsymbol{\alpha}} \Psi(\boldsymbol{x}, \tilde{z})\right)^2}{2\sigma^2}\right) f_{\tilde{Z}}(\tilde{z}) \mathrm{d}\tilde{z}$$

- Numerical integration by 1D quadrature $l(\boldsymbol{c}, \sigma; \boldsymbol{x}, y) \approx \tilde{l}(\boldsymbol{c}, \sigma; \boldsymbol{x}, y)$
- Maximum likelihood to estimate the coefficients

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} \sum_{i=1}^{N} \log \tilde{l}\left(\boldsymbol{c}, \sigma; \boldsymbol{x}^{(i)}, y^{(i)}\right)$$

Cross-validation

- The likelihood is unbounded for $\sigma = 0$: σ is a hyperparameter that can be selected by cross-validation
- The cross-validation score is also used to find a suitable distribution for $\tilde{Z}$ and a truncation scheme

$$\mathcal{A}^{p,q,d} = \left\{\boldsymbol{\alpha} \in \mathbb{N}^d : \|\boldsymbol{\alpha}\|_q \overset{\text{def}}{=} \left(\sum_{i=1}^{d} \alpha_i^q\right)^{\frac{1}{q}} \leq p\right\}$$

Estimation method

Maximum likelihood estimation

- The conditional likelihood for a data point $(\boldsymbol{x}, y)$ is

$$l(\boldsymbol{c}, \sigma; \boldsymbol{x}, y) = \int_{\mathcal{D}_{\tilde{Z}}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y - \sum_{\boldsymbol{\alpha}\in\mathcal{A}} c_{\boldsymbol{\alpha}} \Psi(\boldsymbol{x}, \tilde{z})\right)^2}{2\sigma^2} \right) f_{\tilde{Z}}(\tilde{z}) \mathrm{d}\tilde{z}$$

- Numerical integration by 1D quadrature $l(\boldsymbol{c}, \sigma; \boldsymbol{x}, y) \approx \tilde{l}(\boldsymbol{c}, \sigma; \boldsymbol{x}, y)$
- Maximum likelihood to estimate the coefficients

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} \sum_{i=1}^{N} \log \tilde{l}\left(\boldsymbol{c}, \sigma; \boldsymbol{x}^{(i)}, y^{(i)}\right)$$

Cross-validation

- The likelihood is unbounded for $\sigma = 0$: σ is a hyperparameter that can be selected by cross-validation
- The cross-validation score is also used to find a suitable distribution for $\tilde{Z}$ and a truncation scheme

$$\mathcal{A}^{p,q,d} = \left\{ \boldsymbol{\alpha} \in \mathbb{N}^d : \|\boldsymbol{\alpha}\|_q \stackrel{\text{def}}{=} \left(\sum_{i=1}^d \alpha_i^q\right)^{\frac{1}{q}} \leq p \right\}$$

Outline

Comparisons

Error metric

- The Wasserstein distance of order 2 is the L^2 distance between the quantile functions for continuous random variables:

$$d^2_{\mathrm{WS}}(Y, \hat{Y}) = \|Q_Y - Q_{\hat{Y}}\|^2_{L^2}$$

- Normalized Wasserstein distance

$$\varepsilon = \frac{\mathbb{E}_{\boldsymbol{X}}\left[d^2_{\mathrm{WS}}\left(Y(\boldsymbol{X}), \hat{Y}(\boldsymbol{X})\right)\right]}{\mathrm{Var}\left[Y\right]}$$

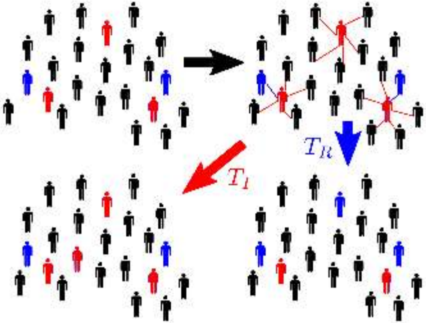
Compared models

- Generalized lambda model (GLaM)
- Stochastic polynomial chaos expansions (SPCE)
- Kernel conditional density estimator (KCDE) Hayfield & Racine (2008) *Nonparametric Econometrics: The np Package*, J. Stat. Softw., 27:1015–1026

Stochastic SIR model in epidemiology

Model description

- $M_t = S_t + I_t + R_t$: total population
- S_t: number of susceptible individuals at time t
- I_t: number of infected individuals at time t
- R_t: number of recovered individuals at time t



Binois et al. (2018) *Practical heteroscedastic Gaussian process modeling for large simulation experiments*, J. Comput. Graph. Stat., 27:808–821

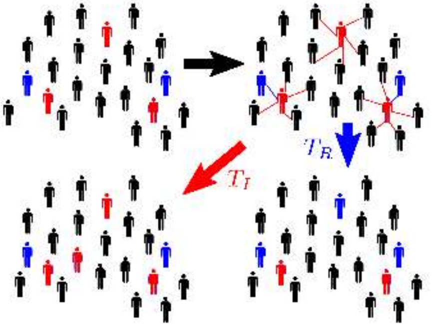
Stochastic SIR model in epidemiology

Model description

- $M_t = S_t + I_t + R_t$: total population
- S_t: number of susceptible individuals at time t
- I_t: number of infected individuals at time t
- R_t: number of recovered individuals at time t

Setup

- Total population $M_t = 2{,}000$
- Initial condition: $S_0 \sim \mathcal{U}(1300, 1800)$, $I_0 \sim \mathcal{U}(20, 200)$
- System dynamics: the contact rate $\beta \sim \mathcal{U}(0.5, 0.75)$, the recovery rate $\gamma \sim \mathcal{U}(0.5, 0.75)$

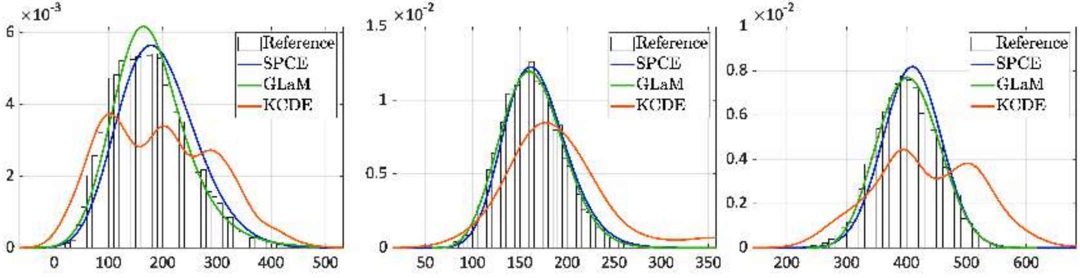


- $Y(\boldsymbol{x})$: total number of infected individuals during the outbreak (without counting I_0)

Binois et al. (2018) *Practical heteroscedastic Gaussian process modeling for large simulation experiments*, J. Comput. Graph. Stat., 27:808–821

PDF predictions

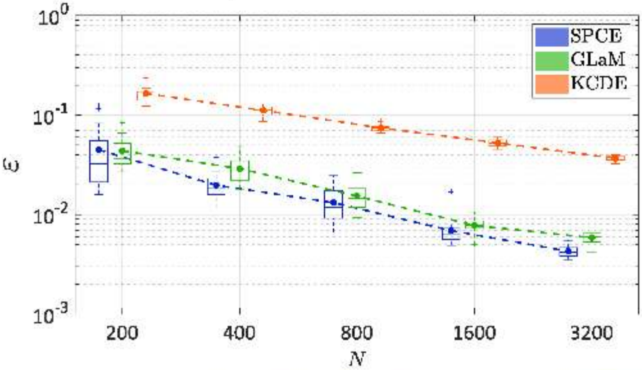
- Surrogates built on an experimental design of size $N = 1{,}600$ generated by the Latin hypercube sampling (without replications)
- 10^4 replications as a reference



$\boldsymbol{x} = (1600, 40, 0.6, 0.55)$ $\boldsymbol{x} = (1700, 120, 0.5, 0.7)$ $\boldsymbol{x} = (1400, 180, 0.7, 0.6)$

Convergence study

- Experimental design of size $N \in \{200; 400; 800; 1{,}600; 3{,}200\}$, no replications
- 20 independent runs for each scenario
- Normalized Wasserstein distance as a performance indicator



Conclusions & Outlook

Conclusions

- Stochastic simulators are used in many fields of applied sciences and engineering
- Building general-purpose emulators is necessary for optimization, sensitivity analysis, etc.
- We propose two surrogate models
  - Generalized lambda models
  - Stochastic polynomial chaos expansions
- Replications are not mandatory ... but can be used

Outlook

- Combinations with other surrogates (e.g., Gaussian processes)
- Sparse techniques, e.g, penalized maximum likelihood estimator $\hat{\boldsymbol{c}} = \arg\min_{\boldsymbol{c}} L(\boldsymbol{c}) + \nu P(\boldsymbol{c})$, e.g., LASSO $P(\boldsymbol{c}) = \|\boldsymbol{c}\|_{l^1}$

Related publications

- X. Zhu and B. Sudret. “Replication-based emulation of the response distribution of stochastic simulators using generalized lambda distributions”. In: *Int. J. Uncertainty Quantification* 10.3 (2020), pp. 249–275. DOI: 10.1615/Int.J.UncertaintyQuantification.2020033029.
- X. Zhu and B. Sudret. “Emulation of stochastic simulators using generalized lambda models”. In: *SIAM/ASA J. Unc. Quant.* (2021). (Submitted). URL: https://arxiv.org/abs/2007.00996.
- X. Zhu and B. Sudret. “Global sensitivity analysis for stochastic simulators based on generalized lambda surrogate models”. In: *Reliab. Eng. Sys. Safety* (2021). (Submitted). URL: https://arxiv.org/abs/2005.01309.
- X. Zhu and B. Sudret. “Stochastic polynomial chaos expansions for emulating stochastic simulators”. In: (2021). (In preparation).

Chair of Risk, Safety & Uncertainty Quantification

www.rsuq.ethz.ch

The Uncertainty Quantification Software

www.uqlab.com

The Uncertainty Quantification Community

www.uqworld.org

Thank you very much for your attention !

Replications

Some results

- Consider a random design of size N/R with replications R, the likelihood is:

$$L(\boldsymbol{c}) = \frac{1}{N} \sum_{i=1}^{N/R} \sum_{r=1}^{R} \log f^{s}_{Y|\boldsymbol{X}} \left(Y^{(i,r)} \middle| \boldsymbol{X}^{(i)}; \boldsymbol{c} \right)$$

- In expectation, we have

$$\mathbb{E}\left[L(\boldsymbol{c})\right] = \mathbb{E}_{\boldsymbol{X},Y}\left[\log f^{s}_{Y|\boldsymbol{X}}\left(Y \mid \boldsymbol{X}; \boldsymbol{c}\right)\right]$$

- The variance of L is given by

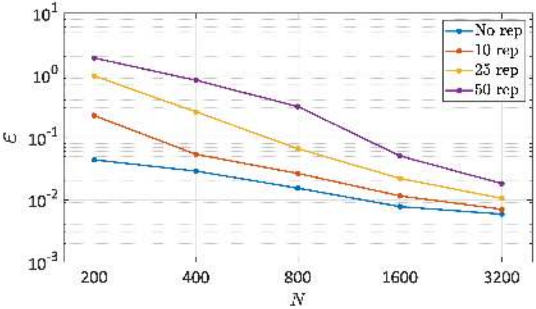
$$\mathrm{Var}\left[L(\boldsymbol{c})\right] = \frac{1}{N}\mathrm{Var}\left[\log f^{s}_{Y|\boldsymbol{X}}\left(Y \mid \boldsymbol{X}; \boldsymbol{c}\right)\right] + \frac{R-1}{N}\mathrm{Var}_{\boldsymbol{X}}\left[\mathbb{E}\left[\log f^{s}_{Y|\boldsymbol{X}}\left(Y \mid \boldsymbol{X}; \boldsymbol{c}\right) \middle| \boldsymbol{X}\right]\right]$$

$R = 1$ (no replications) leads to the minimum variance of $L(\boldsymbol{c})$

Replications (cont.)

Convergence study of the SIR example

- Compare the method that does not need replications with the one based on replications for constructing GLaM
- Replications $R \in \{10; 25; 50\}$
- Total number of model runs $N \in \{200; 400; 800; 1{,}600; 3{,}200\}$



Replications (cont.)

Convergence study of the SIR example

- Compare the method that does not need replications with the one based on replications for constructing GLaM
- Replications $R \in \{10; 25; 50\}$
- Total number of model runs $N \in \{200; 400; 800; 1{,}600; 3{,}200\}$
- Replications are not helpful in this example

However...

- Some methods (e.g., replication-based approaches) rely on the information extracted from replications: trade-off between explorations and replications
- Some methods explore strategies for adaptive designs
- Replications can be used for validations

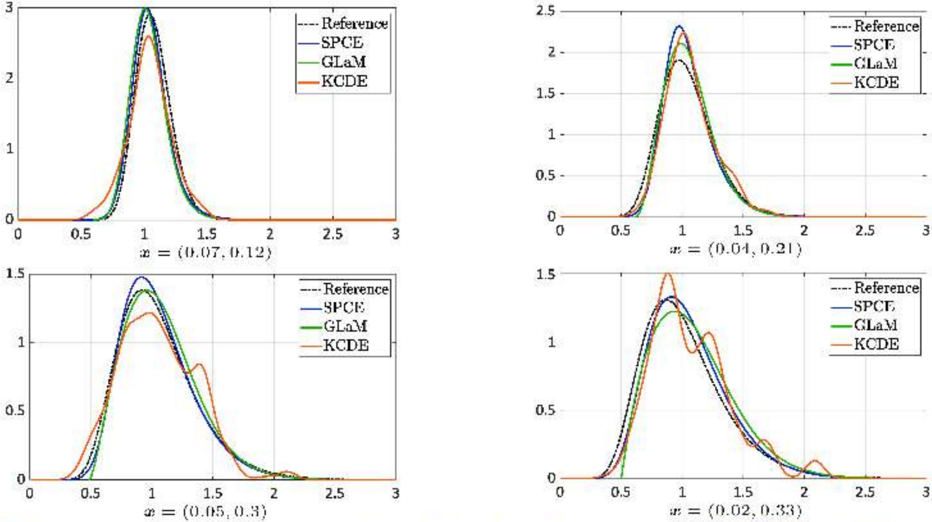
Geometric Brownian motion

$$\mathrm{d}S_t = x_1\, S_t\, \mathrm{d}t + x_2\, S_t\, \mathrm{d}W_t$$

- S_t: price process, W_t: Wiener process, x_1: drift, x_2: volatility
- $X_1 \sim \mathcal{U}(0, 0.1)$, $X_2 \sim \mathcal{U}(0.1, 0.4)$, and $Y(\boldsymbol{x}) = S_1(\boldsymbol{x})$
- The analytical distribution of S_t reads (Itô's calculus):

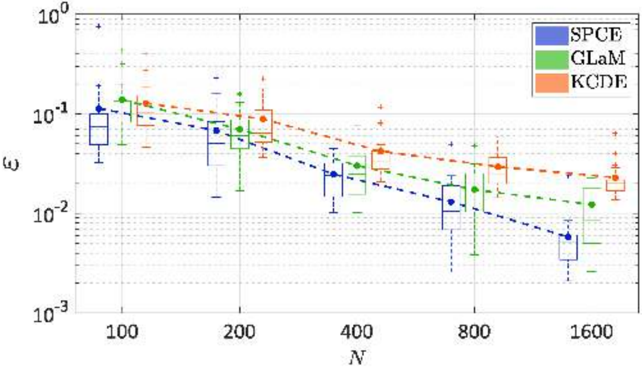
$$S_1(\boldsymbol{x})/S_0 \sim \mathcal{LN}\left(x_1 - \frac{x_2^2}{2}, x_2\right)$$

PDF predictions (ED of size $N = 400$)



Convergence study

- Experimental design of size $N \in \{100; 200; 400; 800; 1{,}600\}$, no replications
- 20 independent runs for each scenario
- Normalized Wasserstein distance as a performance indicator

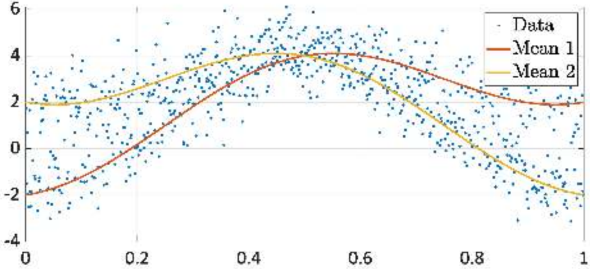


Bimodal toy example

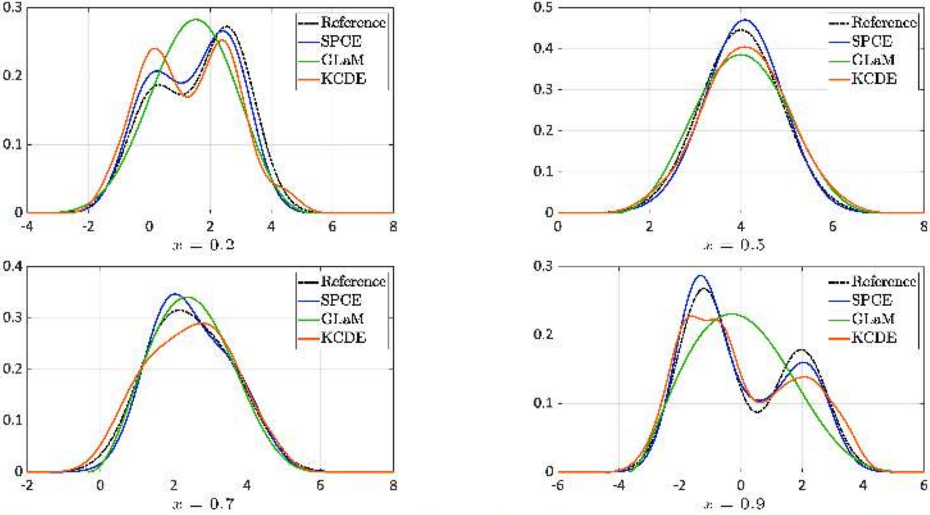
Description of the simulator

$$f_{Y|X}(y \mid X = x) = 0.6\, f_n(4 \sin^2(\pi \cdot x) + 4x - 2) + 0.4\, f_n(4 \sin^2(\pi \cdot x) - 4x + 2)$$

- f_n is the PDF of a normal distribution with mean 0 and standard deviation 0.8, $f_n(t) = \frac{5}{4}\, \varphi\left(\frac{5}{4}t\right)$
- The response distribution is a mixture of Gaussian PDFs
- $X \sim \mathcal{U}(0, 1)$



PDF predictions (ED of size $N = 800$)



Convergence study

- Experimental design of size $N \in \{100; 200; 400; 800; 1{,}600\}$, no replications
- 20 independent runs for each scenario
- Normalized Wasserstein distance as a performance indicator