Emulation of Computer Models with Multivariate Output

Ksenia Kyzyurova
Brown University, Data Science Initiative
ksenia.kyzyurova@gmail.com, ksenia@brown.edu

Providence, Rhode Island, USA
October 17, 2018

Pyroclastic flow

credit: U.S. Geological Survey, volcanoes.usgs.gov

Uncertainty Quantification

Experiments and observations are rare (e.g. volcano eruptions *(Bursik, 2012)*.)

Computer models are simulators based on mathematical representation of reality.

Emulators are fast approximations to computationally expensive simulators *(Sacks, 1989)*.

Titan2D produces height of a pyroclastic flow as an output. This output is not a typical smooth function output of a computer model.

Height values are non-negative and often result in exact zeros (indicating the absence of a flow).

Motivating application

Model *Titan2D*, a model of volcano pyroclastic flow *(Patra, 2015)*.

Inputs: volume of a pyroclastic flow, basal friction angle and initial direction angle.

Output: height of volcano pyroclastic flow.

We are interested in emulation of the maximum height of the flow.

Dominated zero-output

24,576 spatial locations on a grid of 128×192 associated with the island of Montserrat.

500 runs of *Titan2D* at various initial sets of values to the model.

The maximum pile height of the flow is zero for all 500 runs at 8,491 locations.

Spurious (unrealistically small) non-zero numerical values are converted to zeros *(Aghakhani, 2016)*.

Distribution of non-zero height values

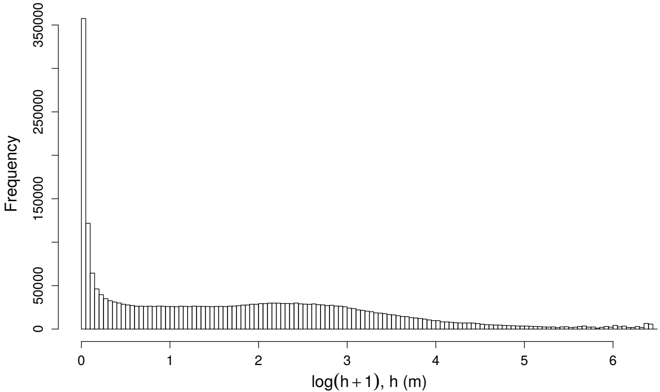


Figure: At the rest of 16,085 locations, about 2/3 of runs resulted in exact zero height values. The distribution of non-zero values is shown.

Zero-inflation problem

is a problem of emulation of *non-negative output* together with

dominated *zero-value output* accompanied by

a *large number of small-height values*.

Censored GASP

Traditional GASP *(Loeppky, 2009)* assumes smooth representation of the output of a computer model.

In this work *a priori* it is known that

> flow height values are non-negative, thus causing inherent restriction on the range of computer model output values,
>
> zero-height output has a *non-zero* probability to occur.

We propose to model height of a pyroclastic flow as *censored at zero traditional GASP of a (sometimes latent) output of a computer model*.

(Wang, 2016; Maatouk, 2017) considered different types of constraints on an underlying function or its derivatives.

Disadvantages of other emulation possibilities

Ignoring zero-output and training emulator only on positive height values.

Emulator needs to perform extrapolation to the zero-output area.

Training emulator on zero- and non-zero- output without discerning between the two.

Leads to the wrong probabilistic assessment of a hazard.

(Spiller, 2014) proposed a combination of the two defined methods to eliminate "non-important" zeros which are far away from non-zero outputs.

Simulation example

For brevity of exposition, assuming univariate input and output,

Data:
Set of n inputs $\mathbf{z}^O = (z_1^O, \ldots, z_n^O)$ with corresponding outputs $(g(z_1^O), \ldots, g(z_n^O))$.

Additionally,

Set of m inputs $\mathbf{z}^C = (z_1^C, \ldots, z_m^C)$ with corresponding outputs $a < g(z_i^C) < b$ for all $i = 1, \ldots, m$.

Simulator

Simulator *Titan2D* produces output such that $a = 0$ and $b = \infty$, i.e.

$$g_0(\cdot) = \max(0, g(\cdot)) \,. \tag{1}$$

Analogously, for any new input z

$$g_0(z) = \begin{cases} g(z), & \text{if } g(z) > 0 \\ 0, & \text{otherwise} \end{cases} . \tag{2}$$

Tobit models (Ertin, 2007; Costa, 2014).

Latent GASP

The latent function $g(\cdot)$ is approximated with GASP *(Bayarri, 2007)*

$$g^M(\cdot) \sim \mathcal{GASP}(\mu(\cdot), \sigma^2 c(\cdot, \cdot)) \quad (3)$$

For any finite number ℓ of input points $\mathbf{z}$

$$g^M(\mathbf{z}) \sim \mathcal{N}(\mu(\mathbf{z}), \sigma^2 \mathbf{C}_z) \quad (4)$$

where $g^M(\mathbf{z}) = \{g^M(\mathbf{z_1}), \ldots, g^M(\mathbf{z}_\ell)\}^{\mathrm{T}} = \{g(\mathbf{z_1}), \ldots, g(\mathbf{z}_\ell)\}^{\mathrm{T}}$ and $\mathbf{C}_z$ is a correlation matrix, for which an element at ith row and jth column is equal to $c(\mathbf{z}_i, \mathbf{z_j}) + \eta \mathbb{I}_{i=j}$, where $\eta \geq 0$ is a parameter which accounts for possible nugget.

Latent GASP joint distribution

Joint distribution of $g^M(\mathbf{z})$ at a set of design points $\mathbf{z} = (\mathbf{z}^O, \mathbf{z}^C)$ may equivalently be written as

$$g^M(\mathbf{z}^O) \sim \mathcal{N}\left(\mu(\mathbf{z}^O), \sigma^2 \mathbf{C}_{z^O}\right), \tag{5}$$

$$g^M(\mathbf{z}^C) \mid g^M(\mathbf{z}^O) \sim \mathcal{N}\left(\mu^*(\mathbf{z}^C), \sigma^{*2}(\mathbf{z}^C)\right), \tag{6}$$

with

$$\mu^*(\mathbf{z}^C) = \mu(\mathbf{z}^C) + c(\mathbf{z}^C, \mathbf{z}^O)\mathbf{C}_{z^O}^{-1}(g_a^M(\mathbf{z}^O) - \mu(\mathbf{z}^O)), \tag{7}$$

$$\sigma^{*2}(\mathbf{z}^C) = \sigma^2(\mathbf{C}_{z^C} - c(\mathbf{z}^C, \mathbf{z}^O)\mathbf{C}_{z^O}^{-1} c(\mathbf{z}^O, \mathbf{z}^C)), \tag{8}$$

where $\mathbf{C}_{z^C}$ and $\mathbf{C}_{z^O}$ are correlation matrices whose (k, l)th elements are given by a correlation function $c(\cdot, \cdot)$.

Censored GASP joint distribution

Corresponding joint distribution of the censored emulator $g_a^M(\mathbf{z})$ of the simulator output $g_a(\mathbf{z})$ at design input points $\mathbf{z}$ is given by

$$g^M(\mathbf{z}^O) \sim \mathcal{N}\left(\mu(\mathbf{z}^O), \sigma^2 \mathbf{C}_{z^O}\right) , \tag{9}$$

$$g_a^M(\mathbf{z}^C) \mid g^M(\mathbf{z}^O) \sim \mathcal{TN}_{(-\infty,a)}\left(\mu^*(\mathbf{z}^C), \sigma^{*2}(\mathbf{z}^C)\right) . \tag{10}$$

Since

$$g_a^M(\mathbf{z}^C) \mid g^M(\mathbf{z}^O) = g^M(\mathbf{z}^C) \mid g^M(\mathbf{z}^O), g^M(\mathbf{z}^C) < \mathbf{a} \, . \tag{11}$$

Predictive distribution of the latent emulator

At any new input $\mathbf{z}'$ latent emulator $g^M(\cdot)$, conditional on evaluations of the computer model $g^M(\mathbf{z})$ at design input points $\mathbf{z}$, is

$$g^M(\mathbf{z}') \mid g^M(\mathbf{z}) \sim \mathcal{N}\left(\mu(\mathbf{z}') + c(\mathbf{z}', \mathbf{z})\mathbf{C}_z^{-1}(g^M(\mathbf{z}) - \mu(\mathbf{z})),\right.$$
$$\left.\sigma^2(1 - c(\mathbf{z}', \mathbf{z})\mathbf{C}_z^{-1}c(\mathbf{z}, \mathbf{z}'))\right). \quad (12)$$

Predictive distribution of a censored emulator

Latent GASP

$$g^M(\mathbf{z}') \mid g^M(\mathbf{z}) = g^M(\mathbf{z}') \mid g^M(\mathbf{z}^O), g^M(\mathbf{z}^C) \tag{13}$$

Since $g^M(\mathbf{z}^C)$ are censored, so that it is known that $g^M(\mathbf{z}^C) < \mathbf{a}$, instead, we are interested in

$$g^M(\mathbf{z}') \mid g^M(\mathbf{z}^O), g^M(\mathbf{z}^C) < \mathbf{a}\,. \tag{14}$$

Predictive distribution

Latent marginal distribution $g^M(\mathbf{z}') \mid g^M(\mathbf{z}) = g^M(\mathbf{z}') \mid g^M(\mathbf{z}^O), g^M_a(\mathbf{z}^C) = g^M(\mathbf{z}') \mid g^M(\mathbf{z}^O), g^M(\mathbf{z}^C) < \mathbf{a}$ is

$$p(g^M(\mathbf{z}') \mid g^M(\mathbf{z}^O), g^M(\mathbf{z}^C) < \mathbf{a}) = \int \cdots \int p(g^M(\mathbf{z}') \mid g^M(\mathbf{z}^O), g^M(\mathbf{z}^C)) \, p(g^M(\mathbf{z}^C) \mid g^M(\mathbf{z}^O), g^M(\mathbf{z}^C) < \mathbf{a}) \, \mathrm{d}g^M(\mathbf{z}^C). \quad (15)$$

Predictive distribution of the emulator $g^M_a(\cdot)$ at a new input to the computer model $\mathbf{z}'$ consists of two parts: a point mass at a and a Lebesgue measure on $\mathbb{R}_{>a}$.

$$g^M_a(\mathbf{z}') \mid g^M(\mathbf{z}^O), g^M(\mathbf{z}^C) < \mathbf{a} = \begin{cases} g^M(\mathbf{z}') \mid g^M(\mathbf{z}^O), g^M(\mathbf{z}^C) < \mathbf{a}, & g^M_a(\mathbf{z}') > a \\ \int_{-\infty}^{a} p(g^M(\mathbf{z}') \mid g^M(\mathbf{z}^O), g^M(\mathbf{z}^C) < \mathbf{a}) \, \mathrm{d}g^M(\mathbf{z}'), & g^M_a(\mathbf{z}') = a \, . \end{cases}$$

Numerical approximation

Distribution (15) is not closed-forrm, but numerical approximation may be obtained.

After getting k samples from the truncated normal distribution $g^M(\mathbf{z}_i)$, conditional on the samples, the following latent distribution may be obtained

$$g^M(\mathbf{z}') \mid g^M(\mathbf{z})_i \sim \mathcal{N}\Big(\mu(\mathbf{z}') + c(\mathbf{z}', \mathbf{z})\mathbf{C}_z^{-1}(g^M(\mathbf{z})_i - \mu(\mathbf{z})), \\ \sigma^2(1 - c(\mathbf{z}', \mathbf{z})\mathbf{C}_z^{-1}c(\mathbf{z}, \mathbf{z}'))\Big) . \quad (16)$$

Simulation example

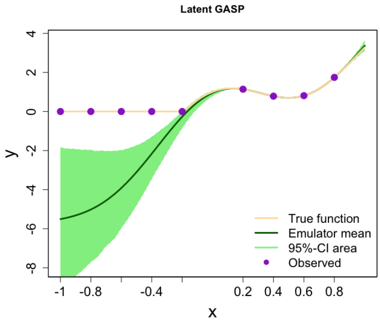
Simulator is given by

$$f_0(x) = \max(0, f(x)), \tag{17}$$

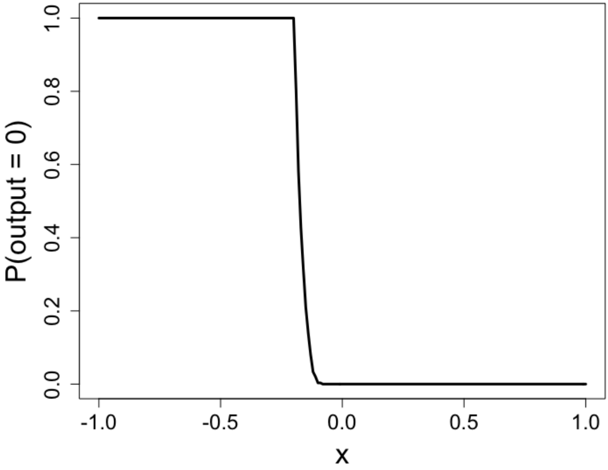
where $f(x) = 3x + \cos(5x)$.

The simulator is observed at
$\mathbf{x}^O = c(0.2, 0.4, 0.6, 0.8)$ and
$\mathbf{x}^C = c(-1, -0.8, -0.6, -0.4, -0.2)$.

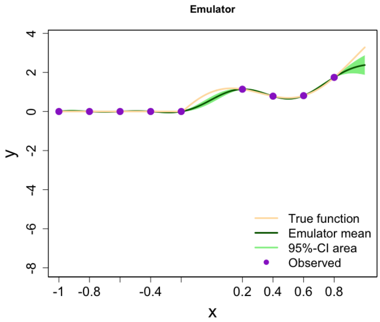
Censored GASP

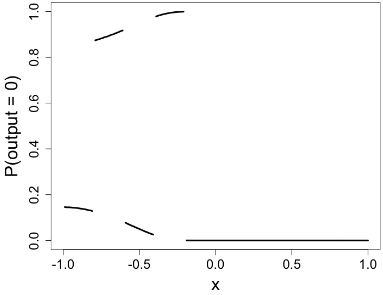


Probability of the zero-height



Projected GASP





Case study results

Table: Comparison of two emulators: projected traditional GASP w/ zeros and censored GASP. Comparison is made on all testing points, including both, zero-height output points and positive-height output points.

GASP	RMSPE	EFC	L_{CI}
Censored	0.649	0.941	1.091
Projected traditional (w/ zeros)	0.618	0.914	1.371

Table: Average probability of a zero-height of three emulators: projected traditional GASP without zeros, projected traditional GASP with zeros and censored GASP at zero-testing points.

GASP	$P(h = 0)$
Censored	0.968
Projected traditional (w/ zeros)	0.504

Pyroclastic flow

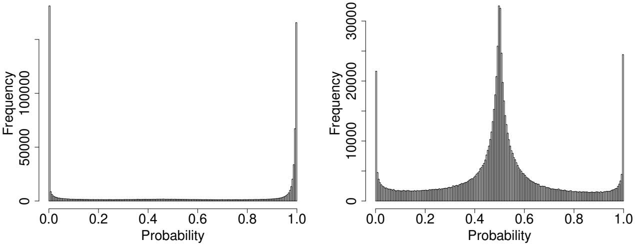


Figure: Left: histogram of probability of zero for censored GASP. These are probabilities for all testing points. Right: histogram of probability of zero for traditional GASP w/ zeros if posterior is projected to be censored at zero.

Conclusion

Censored GASP provides an appropriate emulator for a computer model whose output is inflated with zeros, such that the emulator is adequate for corresponding use for decision support in policy making.

Thank you. Questions.

M. Bursik, *Estimation and propagation of volcanic source parameter uncertainty . . .*, 2012

J. Sacks, *Design and analysis of computer experiments*, 1989

A.P. Patra, *Parallel adaptive numerical simulation of dry avalanches . . .*, 2005

H. Aghakhani, *Heuristic and Eulerian interface capturing approaches for shallow water type flow*, 2014

E. Ertin, *Gaussian process models for censored sensor readings*, 2007

D.R. Costa *Estimation methods for multivariate Tobit confirmatory factor analysis*, 2014

X. Wang, *Estimating Shape Constrained Functions Using a New Class of Gaussian Processes*, 2016

H. Maatouk, *Gaussian process emulators for computer experiments with inequality constraints*, 2017

E.T. Spiller, *Automating emulator construction for geophysical hazard maps*, 2014