On Emulation of Zero-inflated Output of a Computer Model

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

Providence, Rhode Island, USA
October 26, 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: volcano pyroclastic flow height.

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 parameters 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

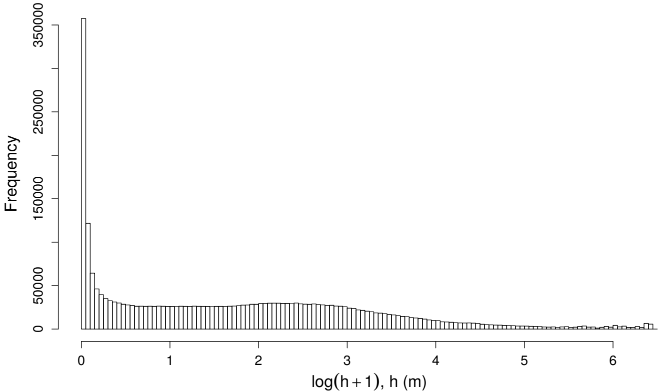


Figure: At the rest of 16,085 locations, about 2/3 of runs resulted in exact zero height values. The distribution of non-zero values is shown.

Zero-inflation problem

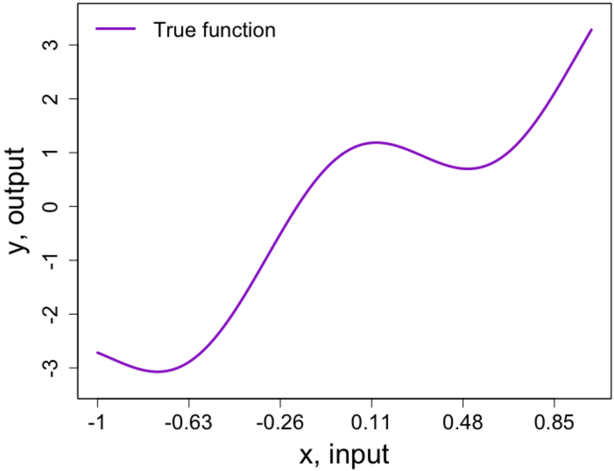
is a problem of emulation of *non-negative output* together with

dominated *zero-value output* accompanied by

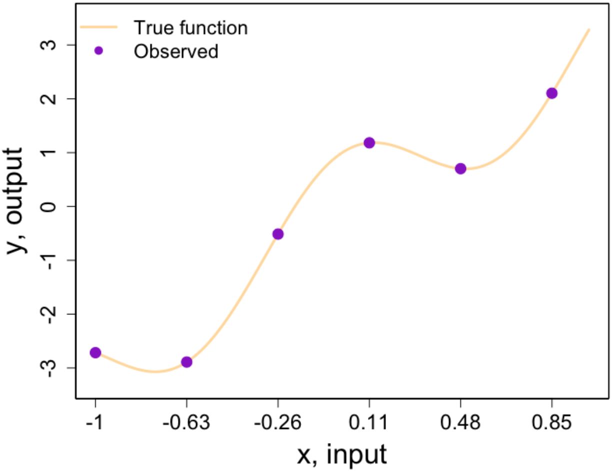
a *large number of small-height values*.

Gaussian process emulator of a computer model

Example

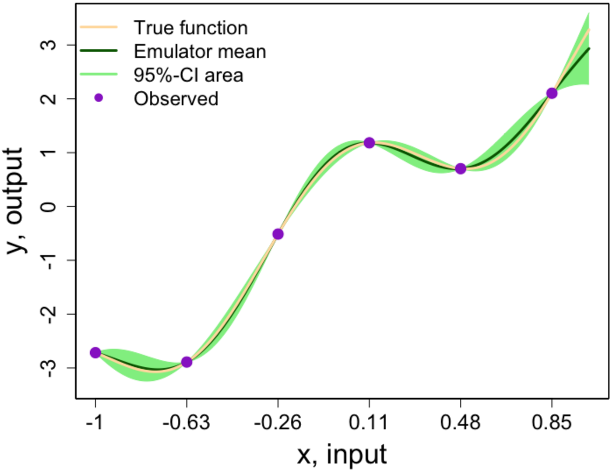


Example



True function
Observed
3
2
1
0
-1
-2
-3
y, output
-1
-0.63
-0.26
0.11
0.48
0.85
x, input

Example



True function
Emulator mean
95%-CI area
Observed
y, output
3
2
1
0
-1
-2
-3
-1
-0.63
-0.26
0.11
0.48
0.85
x, input

Gaussian process emulator

Function g is a simulator of a computer model.
$\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ is a vector of computer model inputs.

If $\{g(\mathbf{x}_1), \ldots, g(\mathbf{x}_m)\} = g(\mathbf{x})$ are the runs of the computer model g at these inputs, then with a Gaussian stochastic process $g^M(\cdot)$ prior on g (*Bayarri, 2007*)

$$g^M(\mathbf{x}) \sim \mathcal{N}(\mu(\mathbf{x}), \sigma^2 \mathbf{K}_x)\,, \tag{1}$$

where $\mu(\mathbf{x}) = (\tilde{\mu}(\mathbf{x}_1), \ldots, \tilde{\mu}(\mathbf{x}_m))$ and $\tilde{\mu}(\cdot)$ is the mean function of the process, σ^2 is the unknown variance and $\mathbf{K}_x$ is the correlation matrix whose (k, l)th element is given by a correlation function $c(\mathbf{x}_k, \mathbf{x}_l)$.

Correlation matrix $\mathbf{C}_z = \mathbf{K}_z + \eta \mathbf{I}$ may be augmented with nugget $\eta = \frac{\tau^2}{\sigma^2}$ (*Gramacy, 2012; Gu, 2018*).

GASP parameters

$\tilde{\mu}(\cdot) = \mathbf{h}(\cdot)^{\mathrm{T}}\beta$ where $\mathbf{h}(\cdot)$ is a vector of regression functions and β are unknown regression coefficients (*Sacks, 1989*)

Correlation function $c(\cdot,\cdot)$ between outputs at two inputs $\mathbf{x}_k$ and $\mathbf{x}_l$ equals

$$c(\mathbf{x}_k, \mathbf{x}_l) = \prod_{j=1}^{d} c(x_{kj}, x_{lj})\,. \tag{2}$$

For the jth coordinate

$$c(x_{kj}, x_{lj}) = \exp\left\{ -\left(\frac{|x_{kj} - x_{lj}|}{\delta_j}\right)^{\alpha_j} \right\}. \tag{3}$$

with range $\delta_j \in (0, \infty)$ and smoothness $\alpha_j \in (0, 2]$.

GASP parameters $\boldsymbol{\theta}_g = (\boldsymbol{\beta}, \sigma^2, \{\alpha_j\}_{j=1,\ldots,m}, \{\delta_j\}_{j=1,\ldots,m}, \eta)$.

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 latent output of a computer model*.

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

Other emulation possibilities

Transformations.

E.g. $\log(h+1)$ *neglects non-zero point mass at zero-height.*

Ignoring zero-output and training only on positive heights.

Demands for 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 last two methods to eliminate "non-important" zeros which are far away from non-zero outputs.

Simulation example

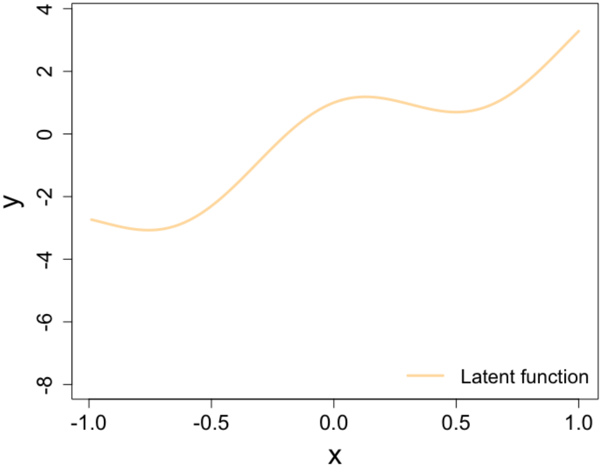
For brevity, assume univariate input and output,

Data consists of two sets of inputs-outputs.

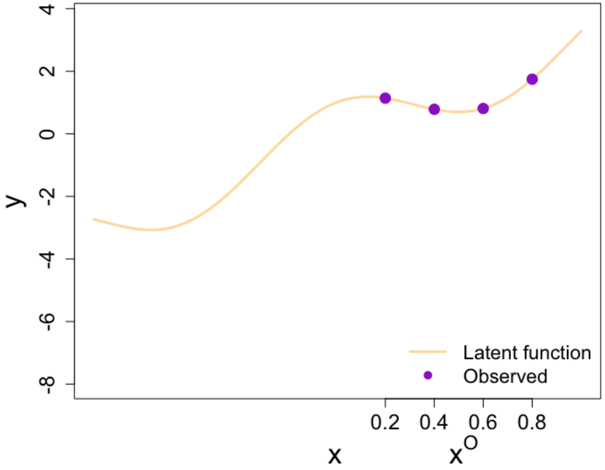
Set of n inputs $\mathbf{x}^O = (x_1^O, \ldots, x_n^O)$ with known outputs $(g(x_1^O), \ldots, g(x_n^O))$.

Set of m inputs $\mathbf{x}^C = (x_1^C, \ldots, x_m^C)$ with corresponding outputs $a < g(x_i^C) < b$ for all $i = 1, \ldots, m$.

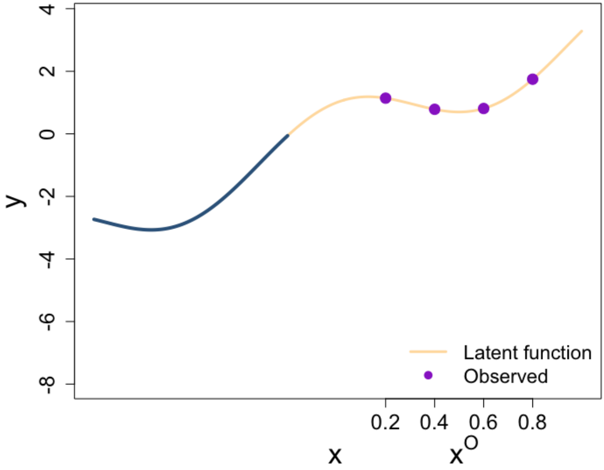
Illustration



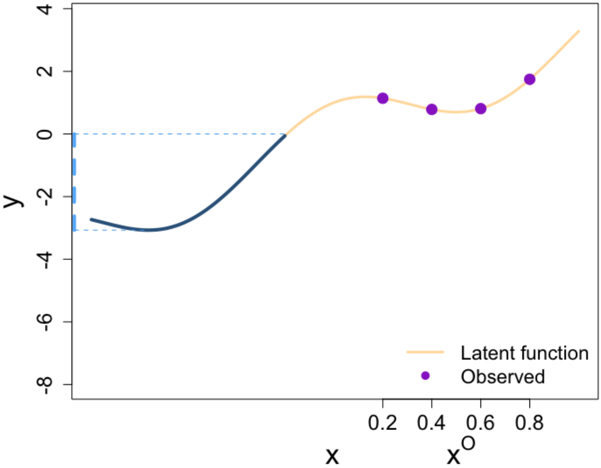
Illustration



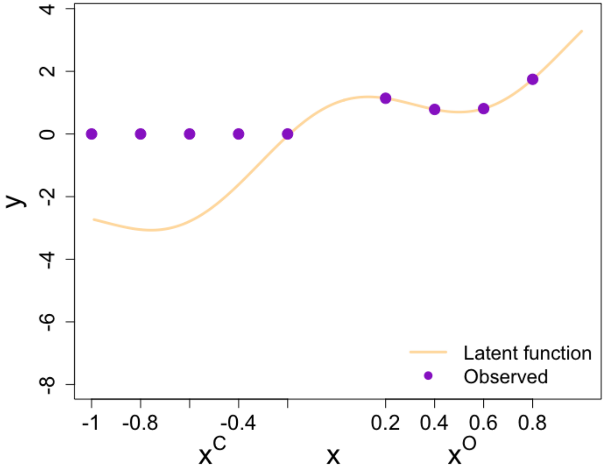
Illustration



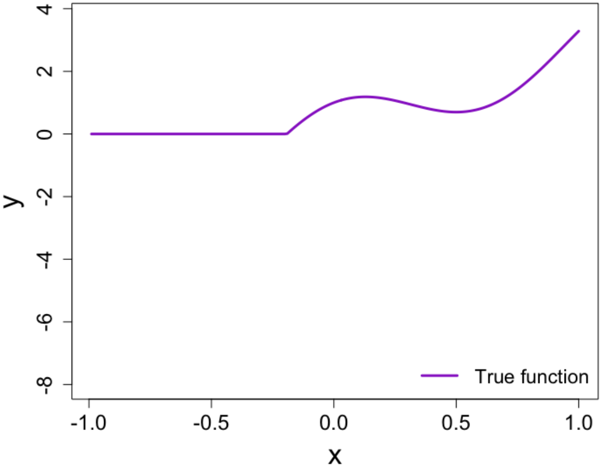
Illustration



Illustration



Illustration



Simulator

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

$$g_0(\cdot) = \max(0, g(\cdot))\,. \tag{4}$$

Analogously, for any new input x

$$g_0(x) = \begin{cases} g(x), & \text{if } g(x) > 0 \\ 0, & \text{otherwise} \end{cases}\,. \tag{5}$$

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

Latent GASP

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

$$g^M(\cdot) \sim \mathcal{GASP}(\mu(\cdot), \sigma^2 c(\cdot,\cdot)) \tag{6}$$

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

$$g^M(\mathbf{x}^O) \sim \mathcal{N}\left(\mu(\mathbf{x}^O), \sigma^2 \mathbf{C}_{x^O}\right) , \tag{7}$$

$$g^M(\mathbf{x}^C) \mid g^M(\mathbf{x}^O) \sim \mathcal{N}\left(\mu^*(\mathbf{x}^C), \sigma^{*2}(\mathbf{x}^C)\right) , \tag{8}$$

with

$$\mu^*(\mathbf{x}^C) = \mu(\mathbf{x}^C) + c(\mathbf{x}^C, \mathbf{x}^O)\mathbf{C}_{x^O}^{-1}(g_a^M(\mathbf{x}^O) - \mu(\mathbf{x}^O)) , \tag{9}$$

$$\sigma^{*2}(\mathbf{x}^C) = \sigma^2(\mathbf{C}_{x^C} - c(\mathbf{x}^C, \mathbf{x}^O)\mathbf{C}_{x^O}^{-1} c(\mathbf{x}^O, \mathbf{x}^C)) , \tag{10}$$

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

Censored GASP

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

$$g^M(\mathbf{x}^O) \sim \mathcal{N}\left(\mu(\mathbf{x}^O), \sigma^2 \mathbf{C}_{x^O}\right), \tag{11}$$

$$g_a^M(\mathbf{x}^C) \mid g^M(\mathbf{x}^O) \sim \mathcal{TN}_{(-\infty,a)}\left(\mu^*(\mathbf{x}^C), \sigma^{*2}(\mathbf{x}^C)\right), \tag{12}$$

because

$$g_a^M(\mathbf{x}^C) \mid g^M(\mathbf{x}^O) = g^M(\mathbf{x}^C) \mid g^M(\mathbf{x}^O), g^M(\mathbf{x}^C) < \mathbf{a}\,. \tag{13}$$

Predictive distribution of the latent emulator

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

$$g^M(\mathbf{x}') \mid g^M(\mathbf{x}) \sim \mathcal{N}\left(\mu(\mathbf{x}') + c(\mathbf{x}', \mathbf{x})\mathbf{C}_x^{-1}(g^M(\mathbf{x}) - \mu(\mathbf{x})),\right.$$
$$\left.\sigma^2(1 - c(\mathbf{x}', \mathbf{x})\mathbf{C}_x^{-1}c(\mathbf{x}, \mathbf{x}'))\right). \quad (14)$$

Predictive distribution of a censored emulator

Latent predictive distribution

$$g^M(\mathbf{x}') \mid g^M(\mathbf{x}) = g^M(\mathbf{x}') \mid g^M(\mathbf{x}^O), g^M(\mathbf{x}^C) \tag{15}$$

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

$$g^M(\mathbf{x}') \mid g^M(\mathbf{x}^O), g^M(\mathbf{x}^C) < \mathbf{a}\,. \tag{16}$$

Predictive distribution

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

$$p(g^M(\mathbf{x}') \mid g^M(\mathbf{x}^O), g^M(\mathbf{x}^C) < \mathbf{a}) = \int \cdots \int p(g^M(\mathbf{x}') \mid g^M(\mathbf{x}^O), g^M(\mathbf{x}^C)) \, p(g^M(\mathbf{x}^C) \mid g^M(\mathbf{x}^O), g^M(\mathbf{x}^C) < \mathbf{a}) \, \mathrm{d}g^M(\mathbf{x}^C). \quad (17)$$

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

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

Numerical approximation

Distribution (17) is not closed-form, but numerical approximation may be obtained.

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

$$g^M(\mathbf{x}') \mid g^M(\mathbf{x})_i \sim \mathcal{N}\Big(\mu(\mathbf{x}') + c(\mathbf{x}', \mathbf{x})\mathbf{C}_x^{-1}(g^M(\mathbf{x})_i - \mu(\mathbf{x})), \\ \sigma^2(1 - c(\mathbf{x}', \mathbf{x})\mathbf{C}_x^{-1}c(\mathbf{x}, \mathbf{x}'))\Big) \ . \quad (18)$$

Simulation example

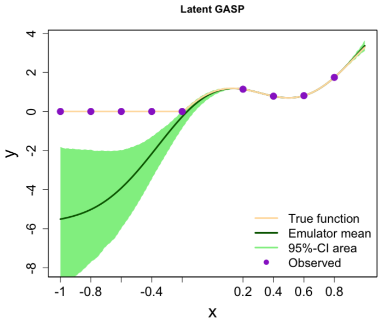
Simulator is given by

$$g_0(x) = \max(0, g(x)), \tag{19}$$

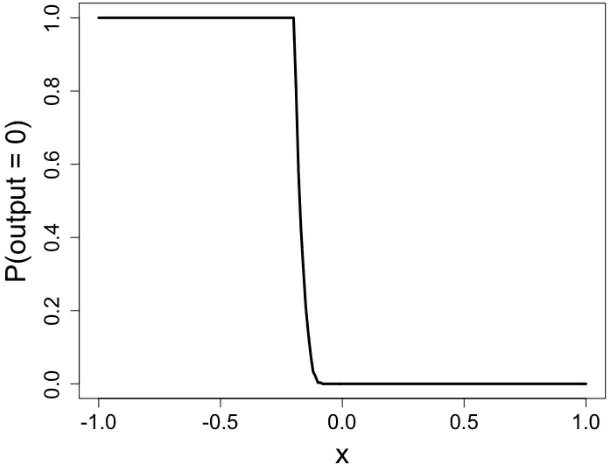
where $g(x) = 3x + \cos(5x)$.

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

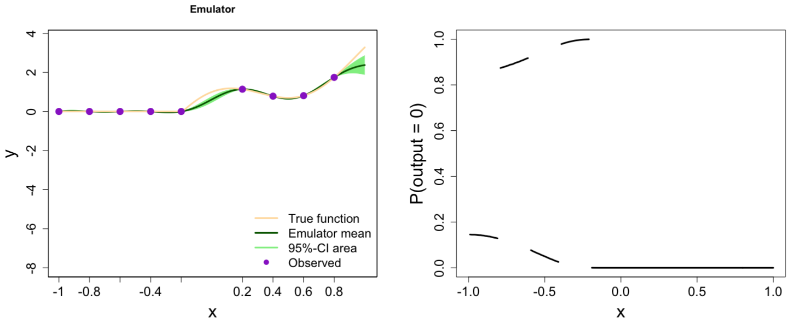
Censored GASP



Probability of the zero-height



Projected GASP



Titan2D application 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

Frequency on hazard probabilities

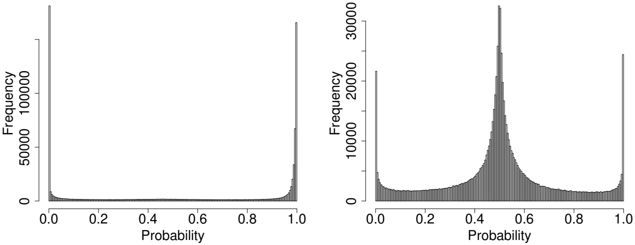


Figure: Left: histogram of probability of zero for censored GASP. These are probabilities for all testing points. Right: histogram of probability of zero for traditional GASP w/ zeros if posterior is projected to be censored at zero.

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