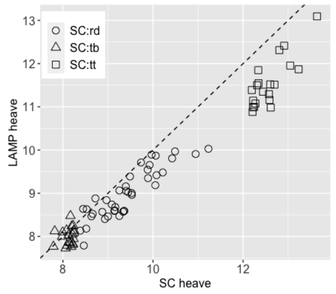
SAMPLING LOW-FIDELITY OUTPUTS FOR ESTIMATION OF HIGH-FIDELITY DENSITY AND ITS TAILS

Minji Kim (UNC-Chapel Hlll), Vladas Pipiras (UNC-Chapel Hlll), Kevin O'Connor (Optiver) and Themistoklis Sapsis (MIT)

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Motivation



Two ship motion programs:
(hi-fi) potential-flow based LAMP and (lo-fi) ODE-based SC.
Generate motions for same random waves determined by record number (random seed).
← Scatter plot of LAMP versus SC heave record maxima including pairs of points corresponding to SC records with 20 largest and 20 smallest observations amongst 2,000.

The goal is to estimate the PDF of LAMP, especially its tails. One possibility is to sample high-dimensional parameters for wave excitation that could lead to extremes, but this is difficult. Another possibility (this work) is to take advantage of "inexpensive" SC that can predict LAMP.

Main Questions

In a multifidelity setting, data are available under the same conditions from two (or more) sources, one being lower-fidelity but computationally cheaper, and the other higher-fidelity and more expensive. This work studies for which low-fidelity (lo-fi) outputs, one should obtain high-fidelity (hi-fi) outputs, if the goal is to estimate the PDF of the latter, especially when it comes to the distribution tails.

Q1: What is an optimal way to sample lo-fi outputs and generate the corresponding hi-fi outputs, so that the estimation of the hi-fi PDF is best? What does optimality mean here? → What p_X should be taken?

Q2: For potential sampling schemes, what are the estimators of the hi-fi PDF? How does one quantify their statistical uncertainty? → How to formulate $\widehat{f}_Y$?

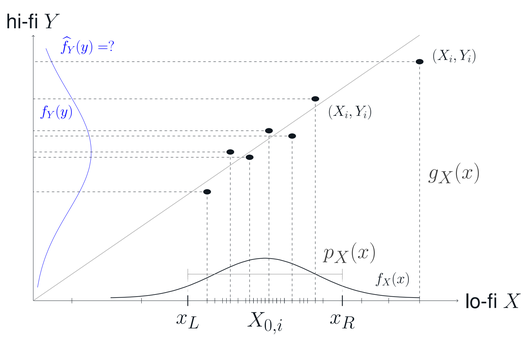
Q3: Should estimation of the PDF be treated separately in the tails, where less (or no) data are available, and how? → Difference between p_X and g_X?

Setting

Preliminaries

$X = X(\omega)$ is lo-fi output and $Y = Y(\omega)$ is hi-fi output for some random seed ω. One can generate X's $(X_{0,1}, \ldots, X_{0,N_0})$ separately from Y's. One can be selective for which obtained X's $(X_1, \ldots, X_N$ with $N \ll N_0)$ to obtain the corresponding values Y $(Y_1, \ldots, Y_N)$.

Visual Illustration



Notation

$f_X(x), f_Y(y)$	PDF of X (sampled at random) and target PDF of Y when $X \sim f_X$
(x_L, x_R)	range of x to estimate f_X well enough to assume it **known**
$X_{0,1}, \ldots, X_{0,N_0}$	data for estimation of f_X
$X_i \in (x_L, x_R), Y_i$	sampled X according to p_X and the corresponding Y-values of X_i
$p_X(x)$	proposal PDF for $x \in (x_L, x_R)$
$g_X(x), g_Y(y)$	proposal PDF for X for the whole range, PDF of Y when $X \sim g_X$
$X_1, \ldots, X_N$	data sampled from g_X, including extremes of $X_{0,i}$
$X_{0,r:N_0}, Y_{r:N}$	rth order statistic of $\{X_{0,i}\}$, $\{Y_i\}$

Importance sampling and estimation

Proposal PDF and importance sampling weights

$$g_X(x) = \begin{cases} c_L\, f_X(x|X \le x_L), & \text{if } x \le x_L, \\ c_0\, p_X(x), & \text{o.w. }, \\ c_R\, f_X(x|X \ge x_R), & \text{if } x \ge x_R, \end{cases} \qquad w(x) = \frac{f_X(x)}{g_X(x)} = \begin{cases} \frac{1}{c_L}\mathbb{P}(X \le x_L), & \text{if } x \le x_L, \\ \frac{1}{c_0}\frac{f_X(x)}{p_X(x)}, & \text{o.w.}, \\ \frac{1}{c_R}\mathbb{P}(X \ge x_R), & \text{if } x \ge x_R. \end{cases}$$

E.g. for $x \ge x_R$, this ensures that all $X_{0,i} \ge x_R$ can be selected in the sampled $X_1, \ldots, X_N$.

Kernel-based estimator of target PDF

$$\widehat{f}_Y(y) = \frac{1}{N}\sum_{i=1}^{N} K_h(y - Y_i)w(X_i),$$

where $h > 0$ is a bandwidth, $K_h(u) = h^{-1}K(h^{-1}u)$ and K is a kernel function.

Modification in the tails

$$\widehat{f}_Y^{(m)}(y) = \begin{cases} g_{\widehat{\xi}_R, \widehat{\delta}_R}(y - y_R), & \text{if } y \ge y_R, \\ \widehat{f}_Y(y), & \text{if } y_L < y < y_R, \\ g_{\widehat{\xi}_L, \widehat{\delta}_L}(-(y - y_L)), & \text{if } y \le y_L, \end{cases}$$

where $g_{\xi,\delta}(u)$ is the PDF of the generalized Pareto distribution (GPD).

Optimality of proposal PDF

We first derive the proposal PDF, p_X, by assuming a noiseless relationship between Y and X, represented as $Y = m(X)$. We suggest applying it to general settings of $Y = m(X) + \epsilon$.

Optimality criteria for the proposal PDF

$$\text{Optimality} : \frac{N\text{Var}(\widehat{f}_Y(y))}{f_Y(y)^2} \simeq \text{const.} \qquad (*)$$

For monotone m, the optimality $(*)$ translates into

$$p_X(x) \propto m'(x), \quad x_L < x < x_R.$$

For piecewise monotone m, the optimality translates into

$$p_X(x) \propto \frac{f_X(x)}{f_Y(m(x))}, \quad x_L < x < x_R,$$

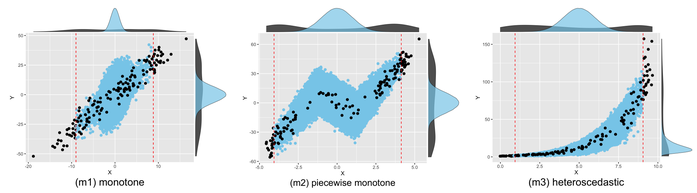
where

$$f_Y(y) = \sum_{j=1}^{n} \frac{f_X(m_j^{-1}(y))}{|m'(m_j^{-1}(y))|}\mathbb{1}_{m(A_j)}(y).$$

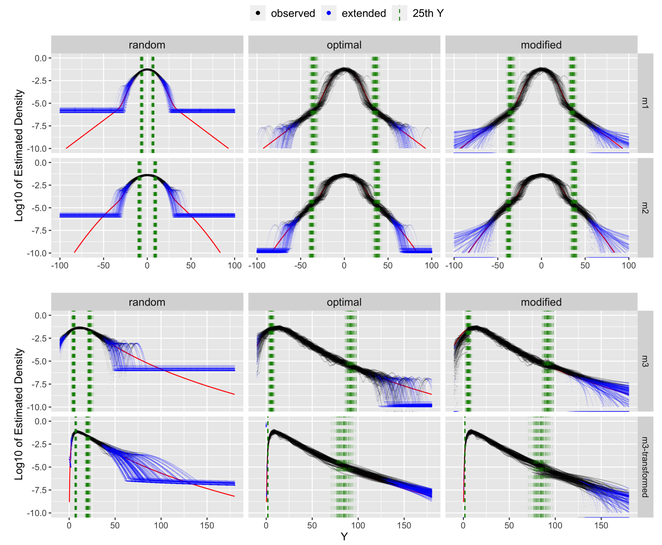
Numerical illustration

$$Y = m_i(X) + \sigma_i(X)\epsilon,\ \epsilon \sim \mathcal{N}(0,1),\ \sigma_i(X) = \begin{cases} \sigma, & \text{if } i = 1, 2, \\ \sigma m_i(x), & \text{if } i = 3. \end{cases}$$

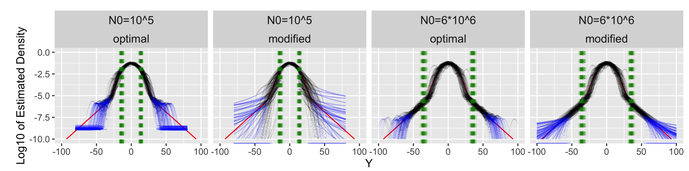
Scatter plot of sampled data for three choices of m



Performance of estimators



Role of N_0 and usefulness of GPD



Optimality of p_X