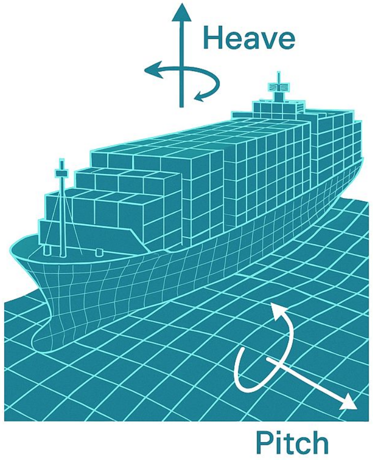
Multi-Fidelity Methods for Distribution Estimation with Focus on Extremes and Naval Applications

2025. 04.04

Ph.D. proposal – Minji Kim

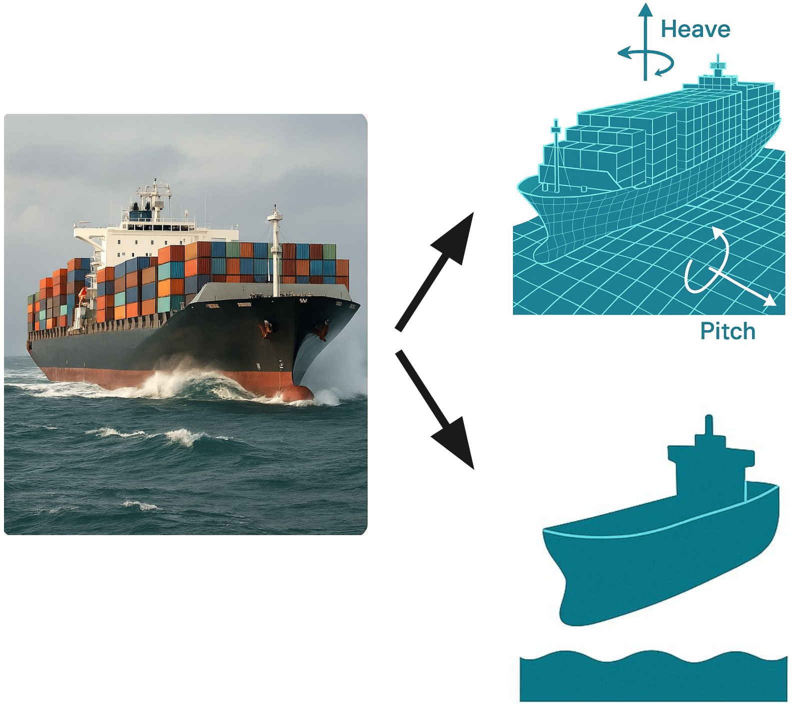
Under the direction of Vladas Pipiras

Modeling physical phenomena using computer simulation codes



- Mathematical models enable simulations as practical alternatives to costly physical experiments
- Simulations help explore extreme conditions and test a wide range of scenarios

Computer simulation codes with different fidelities



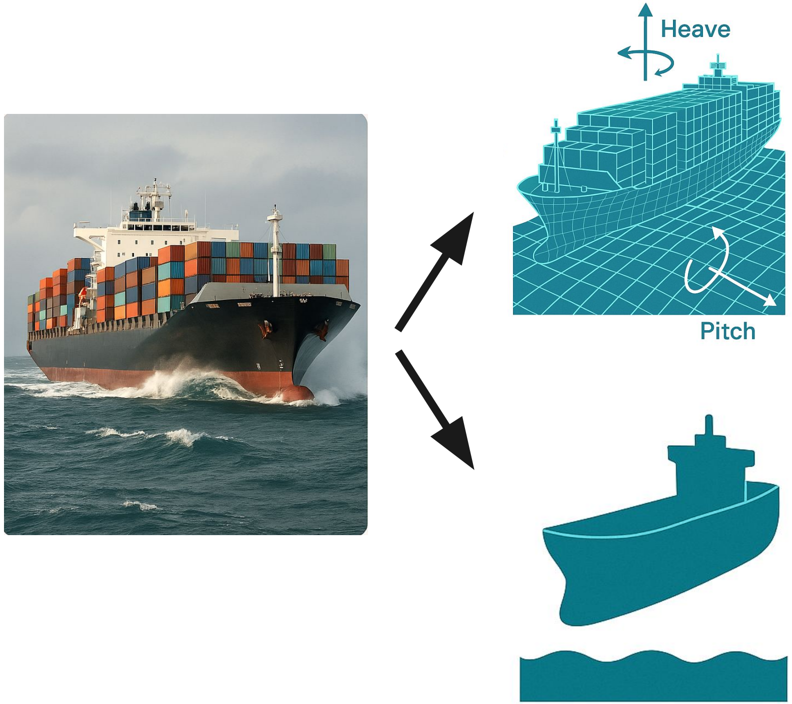
High-Fidelity (LAMP)

- High-fidelity (hi-fi) simulations are accurate but computationally expensive
- Uncertainty quantification (UQ) often requires multiple model evaluations

Low-Fidelity (SC)

- Surrogate (low-fidelity; lo-fi) models approximate behavior with reduced cost
- Fidelity can vary through grid resolution, dimensional reduction, or by simplifying the underlying physical model

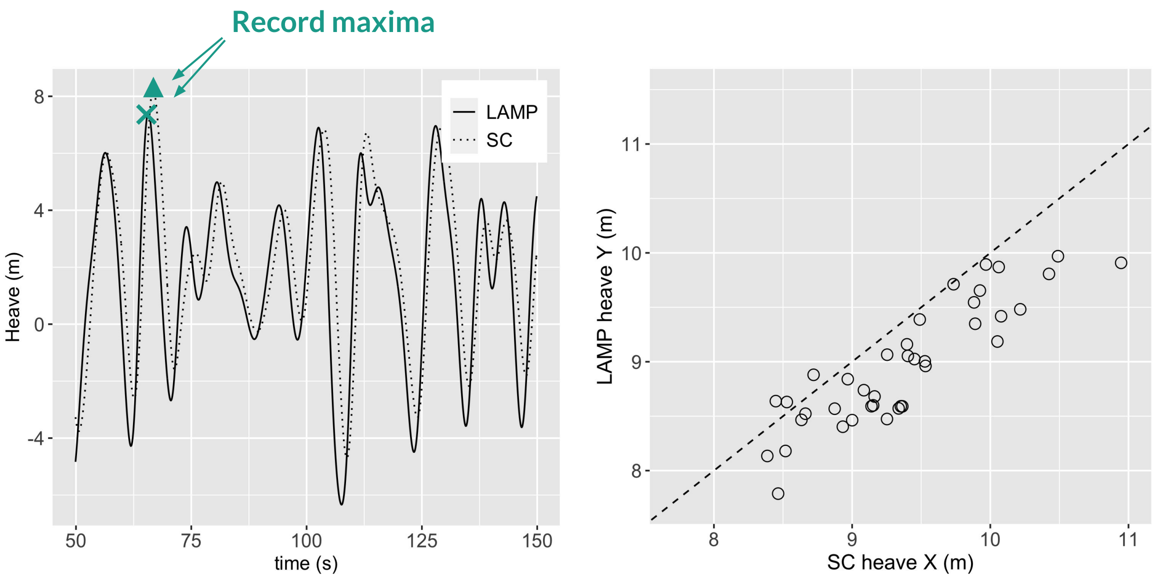
Computer simulation codes with different fidelities



Multi-Fidelity (MF) Methods

- Goal: When multiple models (of different fidelities) are available for the same output quantity, how can we efficiently utilize the data?
- MF approaches aim to leverage lo-fi models to reduce computational costs, while relying on hi-fi outputs to ensure accuracy
- From a statistical perspective, we aim to enhance prediction (with reduced variance) by leveraging abundant low-fidelity outputs

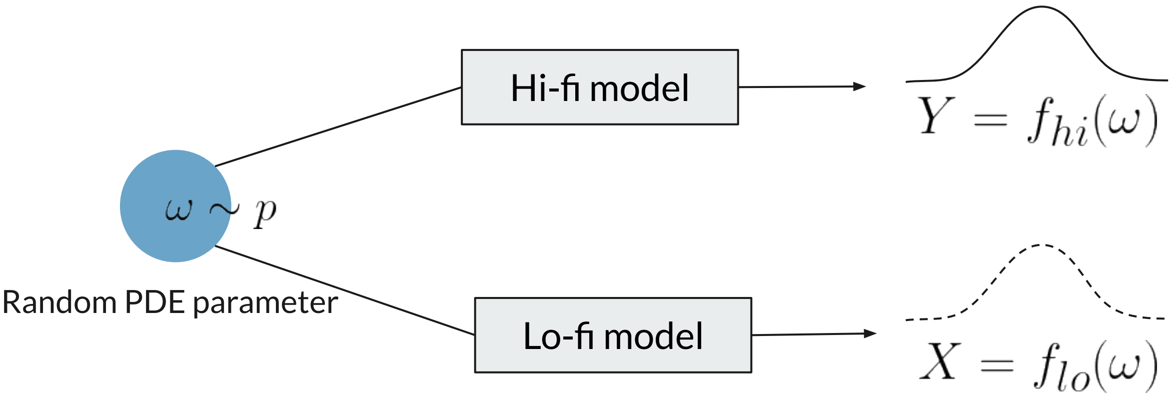
Example observations from High- and Low-fidelity (LAMP and SC) models



(Left) Heave motion for LAMP and SC observed over a 100-second time window. (Right) LAMP versus SC heave record maxima. The dashed line is the 45° line.

Multi-fidelity objectives

Forward Simulation
Uncertainty propagation



Uncertainty
Quantification

Key Question

To better estimate the distribution of high-fidelity outputs, (Goal)

how can we leverage the low- fidelity outputs? (Strategy)

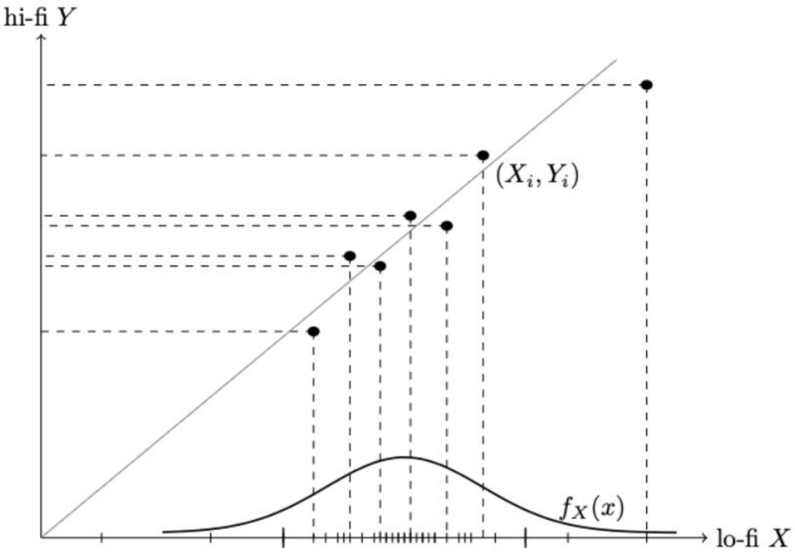
Multi-fidelity strategies

Part1. Selective Sampling

- **low-fidelity model outputs are explored first to determine where to evaluate the high-fidelity model**
- Non-parametric density estimation

Part2. Data Fusion

- **a larger amount of independently obtained low-fidelity data is used to obtain estimators with reduced variance.**
- Parametric estimation
- Non-parametric estimation (future direction)



Selective Sampling

Sampling low-fidelity outputs to estimate high-fidelity density and its tails

Work with Kevin O'Connor, Vladas Pipiras, Themistoklis Sapsis

SIAM/ASA Journal on Uncertainty Quantification ***13****, pp. 30–62, 2025*

Motivation

Quantity of Interest:	$f_Y(y)$
Random sampling:	$(X_1, Y_1), \ \dots (X_n, Y_n),$
Additional data:	available for X

Note: It is possible to generate X and Y separately

- Baseline estimator : $\frac{1}{n}\sum_{i=1}^{n} K_h(y - Y_i)$ Can we do better than this?
- We potentially have more observations available for X, which is correlated with Y. Intuitively, these additional sampling of X should help improve our estimation of the quantity of interest for Y... But how?
- It is inefficient to explore distribution tail relying solely on random sample of high-fidelity model.
- Importance sampling based approach naturally arises in this context.

Existing approach

Quantity of Interest:	$p_a = \mathbb{P}(Y > a)$
Random sampling:	$(X_1, Y_1), \ \ldots (X_n, Y_n),$
Additional data:	available for X

- Baseline estimator : $\hat{p}_a = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(Y_i > a)$ Can we do better than this?

- IS approach aim to bias sample toward the region of interest $\{Y > a\}$. As it requires a large number of sample to estimate the rare event set, multi-fidelity approach instead estimate

$$\{\omega : Y(\omega) > a\} \ \text{with} \ \{\omega : X(\omega) > a\}$$

* Initial ω~p_ω is given

and construct biasing distribution for ω (e.g., mixture of gaussians) based on samples in $\{\omega : X(\omega) > a\}$.

- Sample new parameter ω' from the fitted distribution, evaluate Y's to construct IS estimator.

Methods

- We propose IS-based density estimator, but focus on the direct relationship between X and Y.

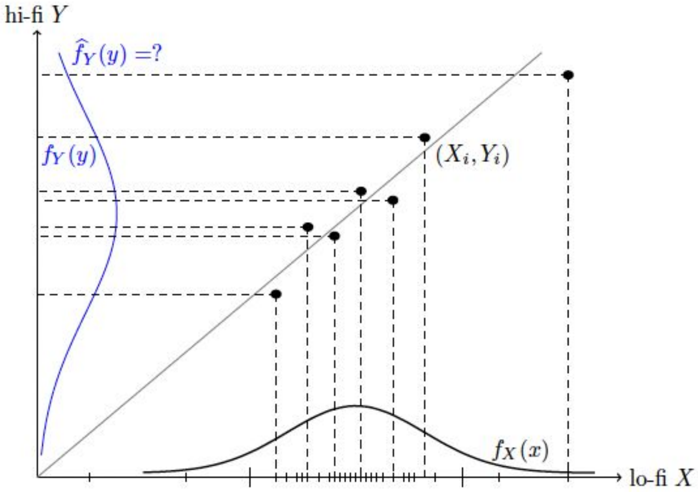
Step 1: generate N_0 samples for X to approximate $X \sim f_X$

Step 2: obtain N samples of X based on the proposal PDF g_X

Step 3: obtain N samples of Y conditionally on sampled Xs'

Step 4: construct the IS estimator as

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



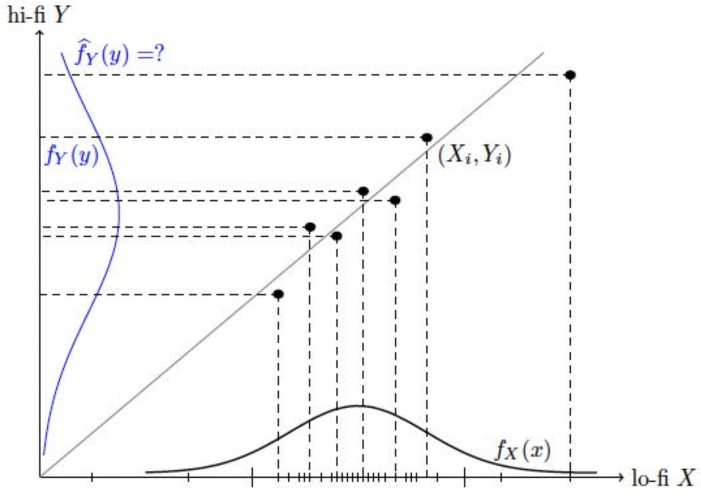
Methods

- Step 1: generate N_0 samples for X to approximate f_X
  One consideration is that we can expect to estimate f_X well over certain range, say, (x_L, x_R).
- We devise the following structure for the proposal PDF:

$$g_X(x) = \begin{cases} c_L f_X(x|X \le x_L) & \text{if} \quad x \le x_L, \\ c_0 p_X(x) \longleftarrow & \text{if} \quad x_L < x < x_R, \\ c_R f_X(x|X \ge x_R) & \text{if} \quad x \ge x_R, \end{cases}$$

On the range (x_L, x_R), we employ importance sampling.
Outside of the range, we ideally sample all extreme outputs.

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



Methods

- Question: What p_X should be taken? In other words, how can we define the **'optimal'** proposal PDF?
- We adopt the following optimality criteria to find optimal p_X :

$$\frac{N\mathrm{Var}(\widehat{f_Y}(y))}{f_Y(y)^2} \simeq \mathrm{const}$$

Remark
If $Y = m(X)$ and m is monotone, the optimality criteria translates to:

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

This suggests that the favored regions for sampling are determined by the rate of change of Y with respect to X.

Methods

- We propose sampling strategy given proposal PDF p_X (Algorithm 1) and the adaptive sampling scheme incorporating m estimation (Algorithm 2)

Algorithm 2 Adaptive Sampling Incorprating m Estimation

Input: PDF f_X, thresholds x_L and x_R

sample (X_i, Y_i) where $X_i \sim \mathrm{Unif}(x_L, x_R)$ for $i = -1, \ldots, -n_0$

construct $D^{(0)} = \{(X_i, Y_i),\ i = -1, \ldots, -n_0\}$

fit piecewise linear regression (PLR) to $D^{(0)}$ to obtain the initial estimate $\widehat{m}^{(0)}$ and its monotone components $\{\widehat{m}_{j,0}, j \in \mathcal{J}^{(0)}\}$

for $t = 1, \ldots, \tilde{N}$ **do**

  $\widehat{f}_{\tilde{Y}}^{(t)}(y) \leftarrow \sum_{j \in \mathcal{J}^{(t-1)}} \frac{f_X(\widehat{m}_{j,t-1}^{-1}(y))}{|\widehat{m}'_{j,t-1}(\widehat{m}_{j,t-1}^{-1}(y))|} \mathbb{1}(y \in \widehat{m}^{(t)}(A_j))$

  $\widehat{p}_X^{(t)}(x) \leftarrow \frac{f_X(x)}{\widehat{f}_{\tilde{Y}}^{(t)}(\widehat{m}^{(t-1)}(x))}$ ▷ construct $\widehat{p}_X$

  normalize $\widehat{p}_X^{(t)}$ on $x_L < x < x_R$

  sample (X_t, Y_t) where $X_t \sim \widehat{p}_X^{(t)}$ ▷ sample new point

  $w(X_t) \leftarrow \frac{f_X(X_t)}{\widehat{p}_X^{(t)}(X_t)}$ ▷ update weights

  update $D^{(t)} = \{(X_{-n_0}, Y_{-n_0}), \ldots, (X_{-1}, Y_{-1}), (X_1, Y_1), \ldots, (X_t, Y_t)\}$

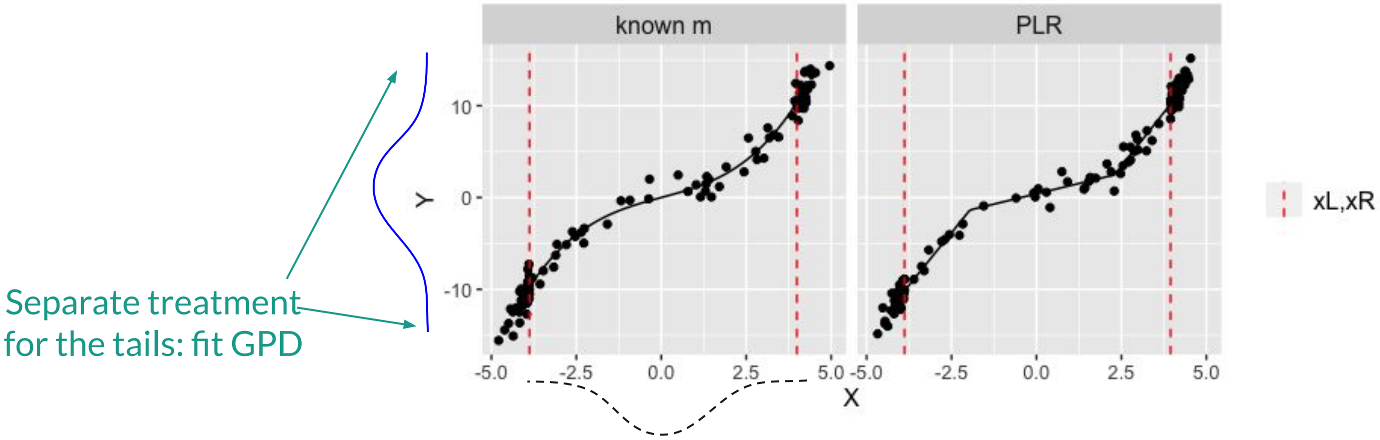
  fit PLR to $D^{(t)}$ to obtain $\widehat{m}^{(t)}$ and its monotone components $\{\widehat{m}_j^{(t)}, j \in \mathcal{J}^{(t)}\}$ ▷ update $\widehat{m}$

end for

Output: Sample $(X_1, Y_1), \ldots, (X_{\tilde{N}}, Y_{\tilde{N}})$.

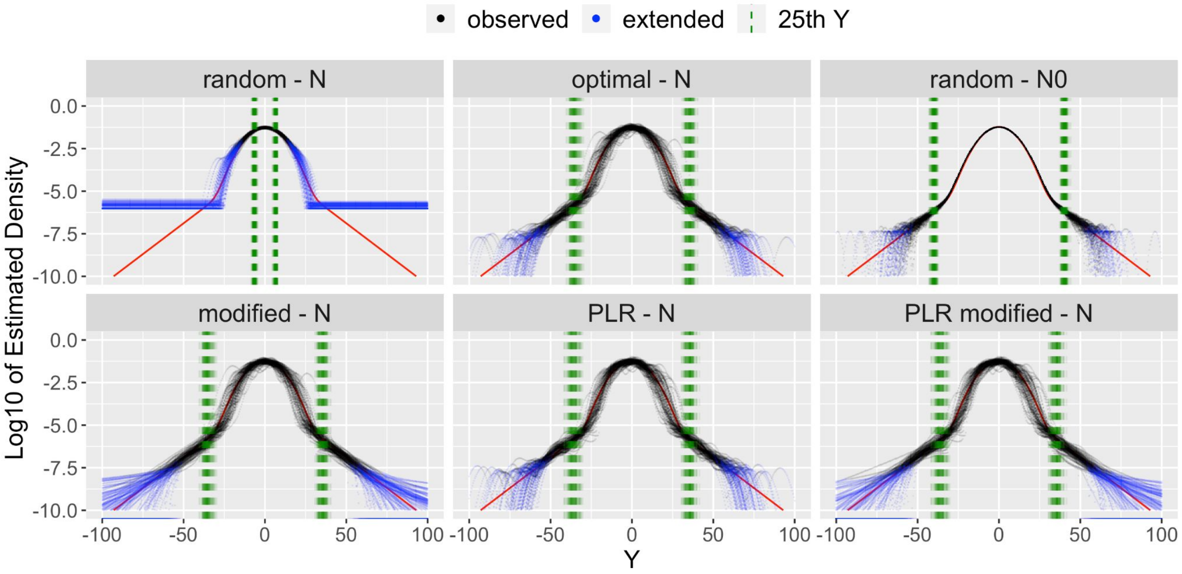
Methods

- We propose sampling strategy given proposal PDF p_X (Algorithm 1) and the adaptive sampling scheme incorporating m estimation (Algorithm 2)



(Left) Sample obtained from the proposal PDF p_X with known m (Algorithm 1) and the true m curve (Right) Sample obtained from the adaptive sampling (Algorithm 2) and the final fitted Piecewise Linear Regression (PLR) curve.

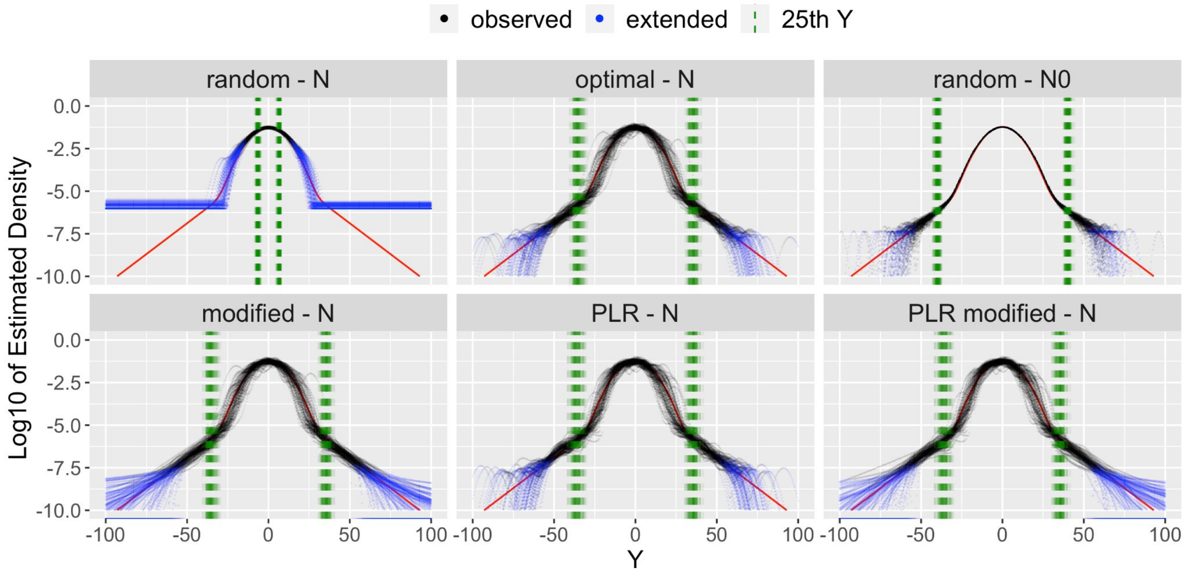
Simulation Result



Estimated versus true log-PDF over 100 realizations for various sampling strategies:

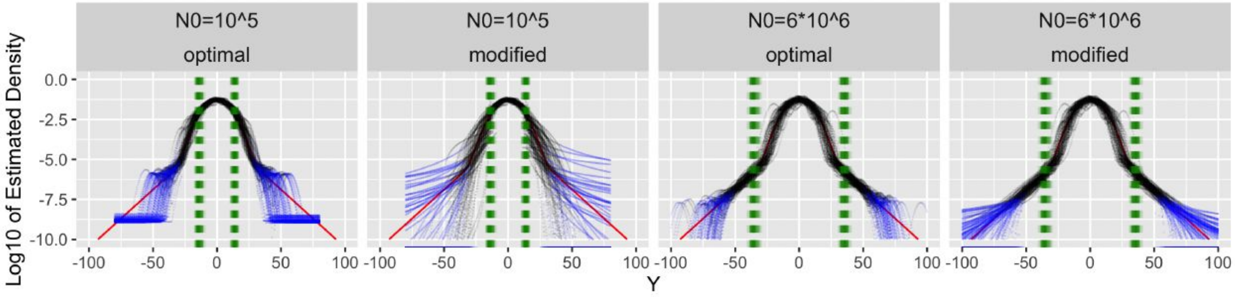
- labels with "- N" or "- N0" refer to the sample size used to compute the estimator; N = 150, N0 = 6*10^6
- "random" represents results from random sampling of Y
- "optimal" and "PLR" show results obtained using the optimal proposal PDF via Algorithm 1 and Algorithm 2, respectively
- any label with "modified" signifies the use of the GPD fit in the tail

Simulation Result



- Using our proposal PDF, we considerably widen the observed sample range and the range where the target PDF is estimated reasonably well.
- In regions with little or no data, the kernel density estimates tend to conform to the shape of the kernel, in our case Gaussian, which is parabolic on the log scale.
- The modified estimator successfully recovers the distribution tail beyond the observed data.

Simulation Result



Comparison of estimated log-PDF across different N0 sizes.

- If GPD fits for too small thresholds, it fails to capture the curvature change in the distribution tails.
- This indicates that GPD fitting with inadequate threshold may not yield **any** benefits, as it does not accurately represent tail behavior

Data Fusion

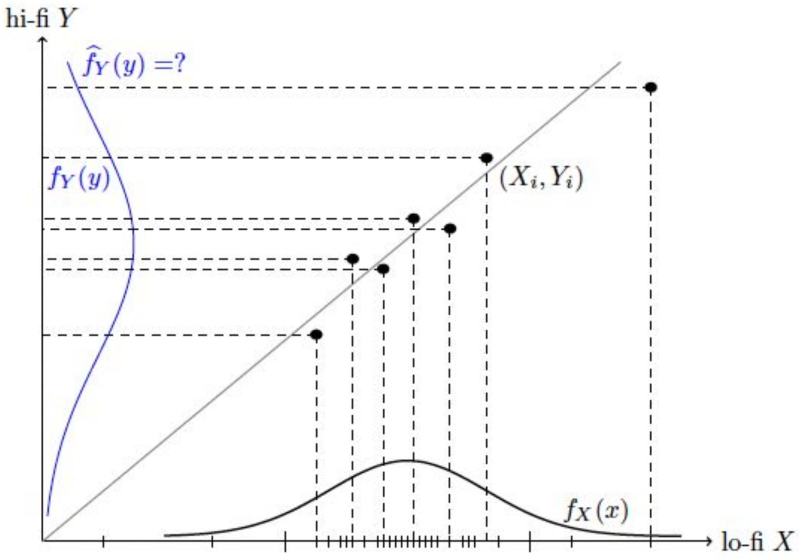
Parametric multi-fidelity Monte Carlo estimation with applications to extremes

Work with Brendan Brown, Vladas Pipiras, revise and resubmit for Technometrics

Multi-fidelity strategies

Part1. Selective Sampling

- In Part 1, we proposed a nonparametric density estimation method using selective sampling.
- This approach assumes the ability to sample Y conditionally on X, which was possible in our motivating application.
- But what if conditional/online sampling is unavailable?



Part2. Data Fusion

- Suppose instead we can generate:
  n joint i.i.d. observations and
  m additional i.i.d. low-fidelity outputs
- We first focus on the **parametric** estimation

Motivation

Quantity of Interest: $\mu = \mathbb{E}Y$

Joint data: $(X_1, Y_1), \ \ldots (X_n, Y_n),$

Additional data: $X_{n+1}, \ \ldots X_{n+m}.$

Note: It is possible to generate X and Y separately

- Baseline estimator: $\bar{Y}_n$

Can we do better than this?

- Now, we have more observations available for X, which is correlated with Y. Intuitively, these additional observations of X should help improve our estimation of the quantity of interest for Y. But how?

- Approximate control variate (ACV) estimator: $\hat{\mu}_{mf} = \bar{Y}_n + \alpha\left(\bar{X}_{n+m} - \bar{X}_n\right),$

$$\mathrm{Var}(\hat{\mu}_{mf})|_{\alpha=\alpha^*} = \frac{\mathrm{Var}(Y)}{n}\left(1 - \frac{m}{m+n}\mathrm{Corr}(X,Y)^2\right), \qquad \alpha^* = \arg\min_{\alpha} \mathrm{Var}(\hat{\mu}_{mf}) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$$

Motivation

Quantity of Interest: $\boldsymbol{\theta}_Y$

Joint data: $(X_1, Y_1), \ \ldots (X_n, Y_n),$

Additional data: $X_{n+1}, \ \ldots X_{n+m}.$

Parametric assumption:

Marginal specification

$$\mathbb{P}(Y \leq y) = F^Y_{\theta_Y}(y), \quad \mathbb{P}(X \leq x) = F^X_{\theta_X}(x),$$

$$\mathbb{P}(X \leq x, Y \leq y) = F_\eta(x, y), \quad \eta = (\theta_X, \theta_Y, \theta_{X,Y})$$

Joint specification

- Baseline estimator: Moment estimator or Maximum Likelihood estimator, using marginal observations of Y.

- Now, we have more observations available for X, which is correlated with Y. Again, these additional observations of X should help improve our estimation of the quantity of interest for Y. But how?

Existing approach

Quantity of Interest: $\mu = \mathbb{E}Y$

Joint data: $(X_1, Y_1),\ \ldots(X_n, Y_n),$

Additional data: $X_{n+1},\ \ldots X_{n+m}.$

1. The previously introduced multi-fidelity estimator is referred to as approximate control variate (**ACV**) type estimator: $\hat{\mu}_{mf} = \bar{Y}_n + \alpha\big(\bar{X}_{n+m} - \bar{X}_n\big)$.
   Extensions considered to multiple low-fidelity models, estimating failure probabilities, CDFs, etc.

2. ACV type estimator can also be driven from the perspective of regression-based **semi-supervised learning** problem, where partially labeled data is used to fit the model.

3. Different approaches from the perspective of semi-supervised learning are available, for example, Chakrabortty and Cai (2018) proposed adaptive imputation strategies to handle **missing labels**.

Methods

Quantity of Interest: $\boldsymbol{\theta}_Y$

Joint data: $(X_1, Y_1), \ \dots (X_n, Y_n),$

Additional data: $X_{n+1}, \ \dots X_{n+m}.$

Parametric assumption:

Marginal specification

$$\mathbb{P}(Y \le y) = F^Y_{\theta_Y}(y), \quad \mathbb{P}(X \le x) = F^X_{\theta_X}(x),$$

$$\mathbb{P}(X \le x, Y \le y) = F_\eta(x, y), \quad \eta = (\theta_X, \theta_Y, \theta_{X,Y})$$

Joint specification

- Baseline estimator

Maximum Likelihood estimator:

$$\hat{\theta}_{Y,bl,ml} = \arg\min_{\theta_Y} \prod_{i=1}^n f_{\theta_Y}(Y_i)$$

Moment estimator:

$$\hat{\theta}_{Y,bl,mom} = g\left(\sum_{i=1}^n h(Y_i)\right)$$

Moment formulation of the parameter: $\theta_Y = g(\mathbb{E}h(Y)), \quad h : \mathbb{R} \to \mathbb{R}^{d_Y}, \ g : \mathbb{R}^{d_Y} \to \mathbb{R}^{d_Y}$

Methods

Observed that the joint ML estimators are effectively the standard MFMC estimators for the first two moments for the Gaussian distribution.

1. Joint maximum likelihood (ML) estimator:

$$(\hat{\theta}_{X,ml}, \hat{\theta}_{Y,ml}, \hat{\theta}_{X,Y,ml}) = \arg\max \prod_{i=1}^{n} f_{(\theta_X,\theta_Y,\theta_{X,Y})}(X_i, Y_i) \underline{\prod_{i=n+1}^{n+m} f_{\theta_X}(X_i)}$$

To assess how this additional term influences the estimation, we analyze a Bivariate Normal case:

Setting: $Y = \alpha + \beta X + \epsilon,\ X \sim N(\mu_X, \sigma_X^2),\ \epsilon \sim N(0, \sigma^2)$

Observe: $\prod_{i=1}^{n} f_{X,Y}(X_i, Y_i) \prod_{j=n+1}^{n+m} f_X(X_i) = \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i) \prod_{j=1}^{n+m} f_X(X_i)$

$f_{Y|X}(y|x) = f_\epsilon(y - \alpha - \beta x)$

Result: $\hat{\mu}_Y = \bar{Y}_n + \hat{\beta}(\bar{X}_{n+m} - \bar{X}_n)$

$$\hat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 + \hat{\beta}^2 \left(\frac{1}{n+m}\sum_{i=1}^{n+m}(X_i - \bar{X}_{n+m})^2 - \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \right)$$

Methods

Motivation: Joint ML estimator requires joint specification of distribution function.
Can we consider ACV type estimators, using only marginal specifications?

2. Moment multi-fidelity estimator:

$$\hat{\theta}_{Y,mom} = g\Big(\overline{h(Y)}_n + \alpha \odot (\overline{h(X)}_{n+m} - \overline{h(X)}_n)\Big)$$

3. Marginal maximum likelihood (MML) multi-fidelity estimator:

$$\hat{\theta}_{Y,mml} = (\hat{\theta}_{X,bl,ml})_n + \beta \odot ((\hat{\theta}_{X,bl,ml})_{n+m} - (\hat{\theta}_{Y,bl,ml})_n)$$

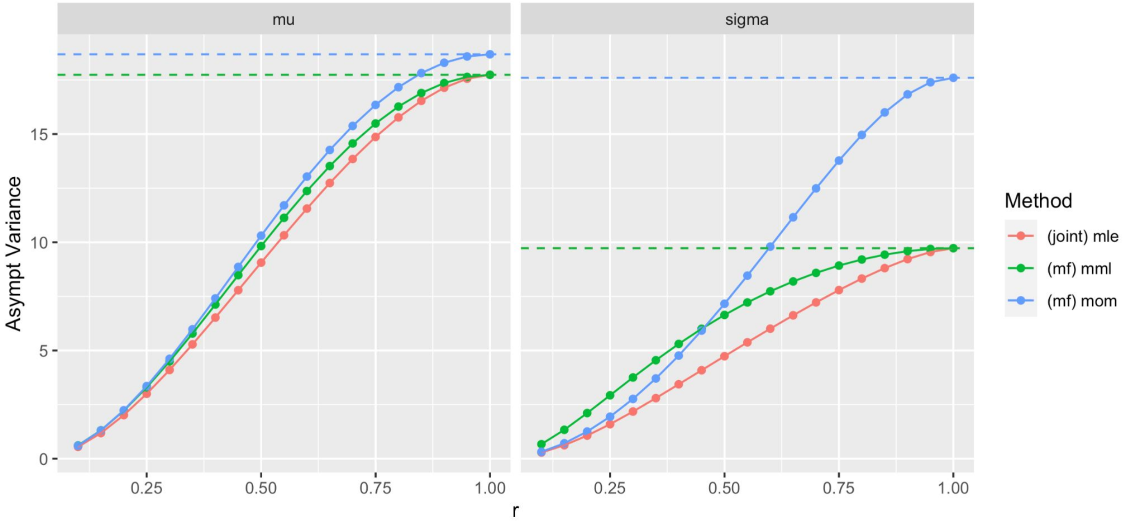
Note: mml estimator resembles a moment estimator formulation, based on the following approximation:

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = I_{\theta^*}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta^*}(X_i) + o_p(1), \quad \dot{\ell}_\theta(x) = (\partial/\partial\theta) \log f_\theta(x) \in \mathbb{R}^{d_Y}$$

Simulation Result

Case study: Bivariate Gumbel Distribution, comparing asymptotic variances of multi-fidelity estimators

$$F_\theta(x) = \exp\{-\exp\{-(x-\mu)/\sigma\}\}$$

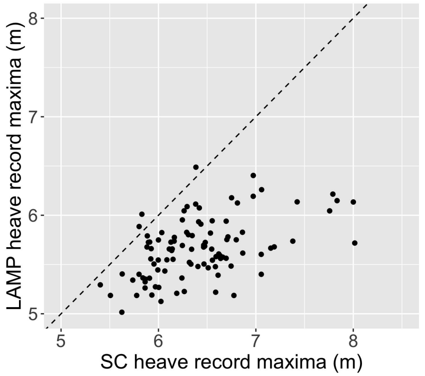


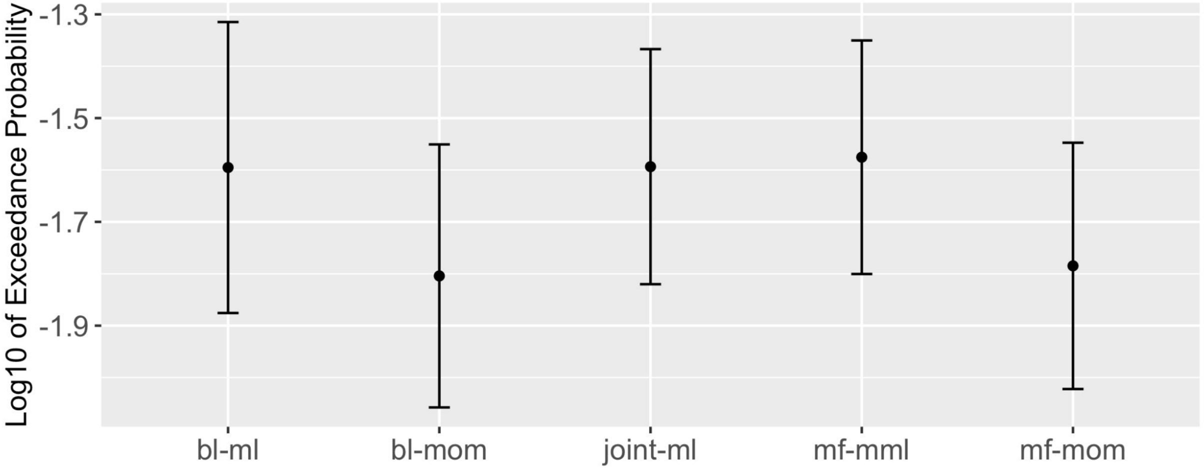
Asymptotic variances of ML (red), multi-fidelity (green and blue, solid), and baseline (green and blue, dashed) estimators for μ and σ across various dependence parameters for the bivariate Gumbel distribution.

Real Data Result

Application to the extreme quantity of interest:

$$q(\mu, \sigma) = \log_{10} \mathbb{P}_\theta(Y > a) = \log_{10}(1 - \exp\{\exp\{-(a - \mu)/\sigma\}\})$$





(Left) Scatterplot of SC and LAMP.

(Right) MFMC estimation result across different methods for estimating an log10 of exceedance probability with a = 6.5

Data Fusion

Future direction (extension to non-parametric approach)

Motivation

Quantity of Interest:	$f_Y(y)$
Joint data:	$(X_1, Y_1),\ \ldots(X_n, Y_n),$
Additional data:	$X_{n+1},\ \ldots X_{n+m}.$

- Baseline estimator : $\frac{1}{n}\sum_{i=1}^{n} K_h(y - Y_i)$ Can we do better than this?

- Available framework:

  According to Owen (2002), $F_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ maximizes $L(F) = \prod_{i=1}^{n} F(\{X_i\}).$

  Empirical distribution function

  Nonparametric likelihood of F

Motivation

Quantity of Interest: $p_{i,j}$

Joint data: $(X_1, Y_1),\ \ldots (X_n, Y_n),$

Additional data: $X_{n+1},\ \ldots X_{n+m}.$

Question: Can we extend the result to the non-parametric case?

- First, focus on the discrete case:

$$p_{i,j} := \mathbb{P}(Y = a_i, X = b_j),\ Y \in \{a_1, \ldots, a_I\},\ X \in \{b_1, \ldots, b_J\}$$

- The joint ML estimator gives us:

$$\hat{p}_{i,j,ml} = \frac{\sum_{k=1}^{n} \mathbb{I}(Y_k = a_i, X_k = b_j)}{\sum_{k=1}^{n} \mathbb{I}(X_k = b_j)} \left(\frac{1}{n+m} \sum_{k=1}^{n+m} \mathbb{I}(X_k = b_j) \right)$$

- We showed that th is also equivalent to the multivariate ACV type multi-fidelity estimator.

Thank you!