This paper is available here →

A Study on Intentional-Value-Substitution Training for Regression with Incomplete Information

○ [1] **Takuya Fukushima**, [1] Tomoharu Nakashima, [1*]Taku Hasegawa, [2**]Vicenç Torra

[1] Osaka Prefecture University [2] Hamilton Institute, Maynooth University

[*] Current affiliation: NTT Media Intelligence Laboratory, NTT Corporation

[**] Current affiliation: Umeå University

Goal

To obtain a robust model for a dataset that contains missing values only at test phase

How can we use the complete information in a training dataset to address missing values during test phase?

Assumptions

- Regression problems
- A training dataset is complete
- A test dataset contains missing values
- The missing probability in the test dataset is unknown
- Which features are possibly deficit is known

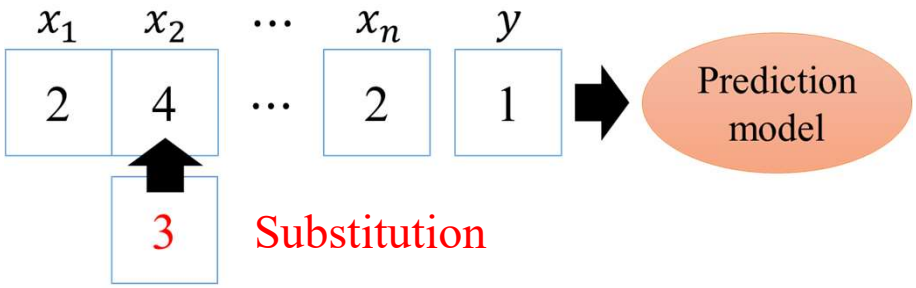
Idea

Intentional-Value-Substitution Training

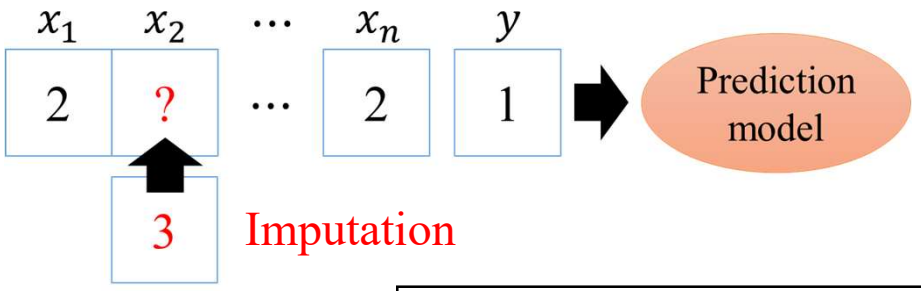
- Consider missing during a model-training phase
- Minimize the error between true function and model's output without predicting missing values

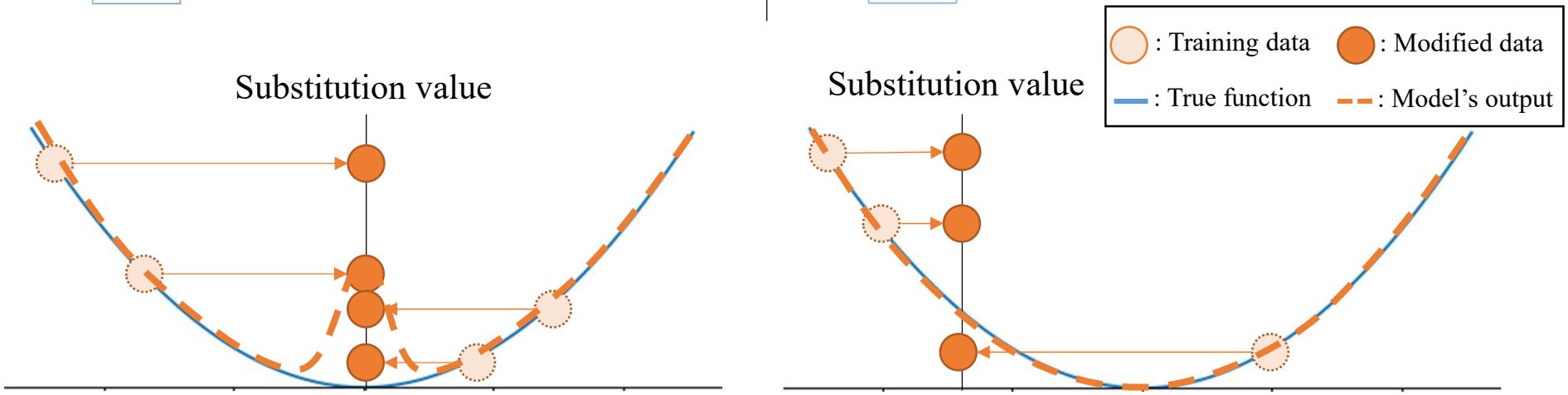
Intentional-Value-Substitution Training

Training phase



Test phase





Optimal substitution value

$$\psi'_{\boldsymbol{x}_{\mathrm{mis}}}(\boldsymbol{x}_{\mathrm{obs}}) = \arg\min_{\boldsymbol{x}'_{\mathrm{mis}}} \{ \int_{D_{\mathrm{mis}}} p(\boldsymbol{x}_{\mathrm{mis}}|\boldsymbol{x}_{\mathrm{obs}}) f(\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}_{\mathrm{mis}})\, dX_{\mathrm{mis}} - f(\boldsymbol{x}_{\mathrm{obs}}, \boldsymbol{x}'_{\mathrm{mis}}) \}^2$$

Proposal

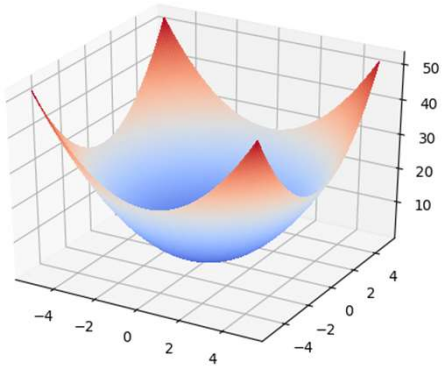
IVS Training

with single missing (previous method) → with multiple missing

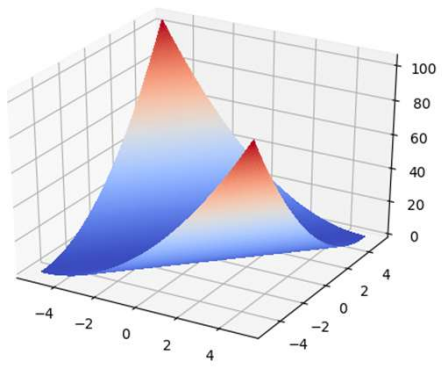
A Study on Intentional-Value-Substitution Training for Regression with Incomplete Information

Experiments

- Function approximation problems
- Pairwise independent
- Drawn from uniform distributions
- $d = 3$
- Substitution probability
- $P_{sub} \in \{0.00, 0.25, 0.50, 0.75, 0.90\}$



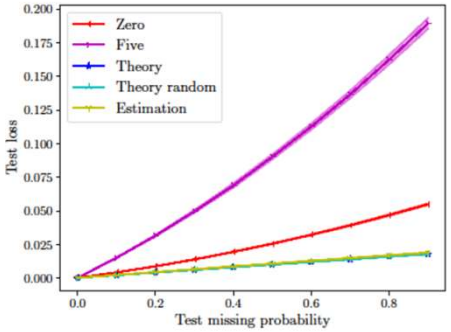
$$f_1(\boldsymbol{x}) = \sum_{k}^{n} x_k^2, \quad (-5 < x_k < 5)$$



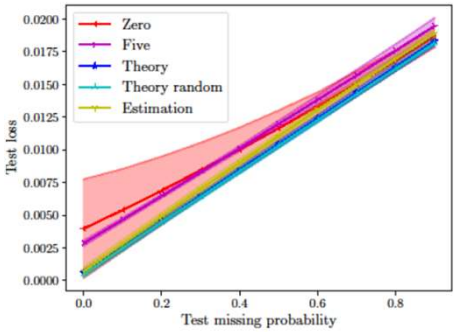
$$f_2(\boldsymbol{x}) = (x_1 - \sum_{k=2}^{d} x_k)^2, \quad (-5 < x_k < 5)$$

Results (excerpt)

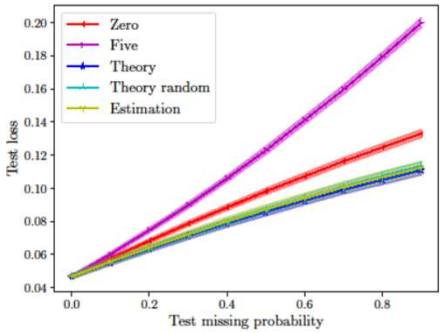
x-axis: P_{mis} y-axis: test loss



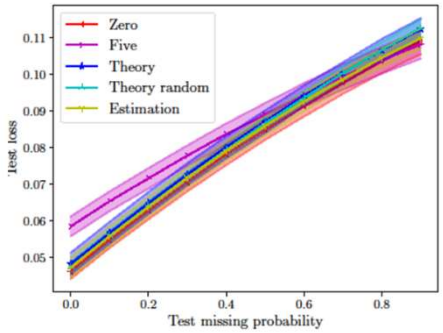
P_{sub} on f_1: 0.0



P_{sub} on f_1: 0.50



P_{sub} on f_2: 0.0



P_{sub} on f_2: 0.50

Conclusion

- Proposed the estimation method of the optimal substitution value in IVS training
- Shown that the validity of the robust model against the loss for unknown data that contain multiple missing by estimating the optimal substitution values