# Input Uncertainty Attribution by Uncertainty Propagation

*Benedikt Kantz* [1] *Sophie Steger* [1] *Clemens Staudinger* [2] *Christoph Feilmayr* [2] *Johann Wachlmayr* [3] *Alexander Haberl* [2] *Stefan Schuster* [2] *Franz Pernkopf* [1 4]

7.4.2025
Presented at the IEEE ICASSP 2025

Contact: benedikt.kantz@tugraz.at
[1]Signal Processing and Speech Communication Laboratory, Technical University Graz, Graz, Austria [2]voestalpine Stahl GmbH, Linz, Austria [3]K1-MET GmbH, Linz, Austria [4]Christian Doppler Laboratory for Dependable Intelligent Systems in Harsh Environments, Graz, Austria

SCIENCE
PASSION
TECHNOLOGY

voestalpine
ONE STEP AHEAD.

# Motivation

- Industrial Application + machine learning (ML) – require accurate & explainable models [1]
- Uncertainty modelling enables another dimension of transparency [2] & improves human-machine decision-making [3], [4]

# Motivation

- Industrial Application + ML – require accurate & explainable models [1]
- Uncertainty modelling enables another dimension of transparency [2] & improves human-machine decision-making [3], [4]

➔ join explainable artificial intelligence (XAI) & uncertainty quantification (UCQ) to attribute uncertainties!

# Motivation

- Industrial Application + ML – require accurate & explainable models [1]
- Uncertainty modelling enables another dimension of transparency [2] & improves human-machine decision-making [3], [4]

➔ join XAI & UCQ to attribute uncertainties!

➔ Existing methods [5]–[7] do not consider ground truths & Uncertainty propagation [8]

# Uncertainty propagation [8]

## Linear uncertainty propagation

Linear function:

$$\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{1}$$

Forward propagation [8]:

$$\boldsymbol{\Sigma}_{\mathbf{g}} = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{A}^{T}. \tag{2}$$

$\mathbf{x} \in \mathbb{R}^{d}$ ... data sample with $d$ features

$\mathbf{g}(\mathbf{x}) \in \mathbb{R}^{o}$ ... ground truth function with $o$ features

$\mathbf{A} \in \mathbb{R}^{o \times d}$ ... affine parameter

$\boldsymbol{\Sigma}_{\mathbf{x}} \in \mathbb{R}^{d \times d}$ ... input covariance matrix

$\boldsymbol{\Sigma}_{\mathbf{g}} \in \mathbb{R}^{o \times o}$ ... output covariance matrix

# Uncertainty propagation [8]

## Nonlinear uncertainty propagation

Taylor series expansion:

$$\mathbf{g}(\mathbf{x}') = \mathbf{g}(\mathbf{x}) + \mathbf{J}_{\mathbf{x}}^{T}(\mathbf{x}' - \mathbf{x}) + \mathcal{O}(\mathbf{x}' - \mathbf{x}), \tag{3}$$

Forward propagation:

$$\boldsymbol{\Sigma}_{\mathbf{g}} = \mathbf{J}_{\mathbf{x}}^{T}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{J}_{\mathbf{x}}. \tag{4}$$

# Uncertainty propagation [8]

## Nonlinear uncertainty propagation

Taylor series expansion:

$$\mathbf{g}(\mathbf{x}') = \mathbf{g}(\mathbf{x}) + \mathbf{J}_{\mathbf{x}}^{T}(\mathbf{x}' - \mathbf{x}) + \mathcal{O}(\mathbf{x}' - \mathbf{x}), \tag{3}$$

Forward propagation:

$$\boldsymbol{\Sigma}_{\mathbf{g}} = \mathbf{J}_{\mathbf{x}}^{T}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{J}_{\mathbf{x}}. \tag{4}$$

Assuming a single-output function $g(\boldsymbol{x})$ and constrain $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ to an *uncorrelated* uncertainty $\mathrm{diag}(\boldsymbol{\sigma}_{\mathbf{x}}^2)$, the expression simplifies to

$$\sigma_{g(\mathbf{x})}^2 = \nabla_{\mathbf{x}} g(\mathbf{x})^T \mathrm{diag}(\boldsymbol{\sigma}_{\mathbf{x}}^2) \nabla_{\mathbf{x}} g(\mathbf{x})$$
$$\sigma_{g(\mathbf{x})}^2 = (\nabla_{\mathbf{x}} g(\mathbf{x})^2)^T \boldsymbol{\sigma}_{\mathbf{x}}^2$$

# Uncertainty propagation [8]

## Nonlinear uncertainty propagation

Taylor series expansion:

$$\mathbf{g}(\mathbf{x}') = \mathbf{g}(\mathbf{x}) + \mathbf{J}_{\mathbf{x}}^{T}(\mathbf{x}' - \mathbf{x}) + \mathcal{O}(\mathbf{x}' - \mathbf{x}), \tag{3}$$

Forward propagation:

$$\boldsymbol{\Sigma}_{\mathbf{g}} = \mathbf{J}_{\mathbf{x}}^{T}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{J}_{\mathbf{x}}. \tag{4}$$

Assuming a single-output function $g(\boldsymbol{x})$ and constrain $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ to an *uncorrelated* uncertainty $\mathrm{diag}(\boldsymbol{\sigma}_{\mathbf{x}}^2)$, the expression simplifies to

$$\sigma_{g(\mathbf{x})}^2 = \nabla_{\mathbf{x}} g(\mathbf{x})^T \mathrm{diag}(\boldsymbol{\sigma}_{\mathbf{x}}^2) \nabla_{\mathbf{x}} g(\mathbf{x})$$

$$\underbrace{\sigma_{g(\mathbf{x})}^2}_{\checkmark} = \underbrace{(\nabla_{\mathbf{x}} g(\mathbf{x})^2)^T}_{\checkmark} \underbrace{\boldsymbol{\sigma}_{\mathbf{x}}^2}_{?}$$

# Smoothness constrained attribution (SCA) I

## Problem proposal

Minimization of error:

$$\boldsymbol{\sigma}_{\mathbf{x}}^{2*} = \arg\min_{\boldsymbol{\sigma}_{\mathbf{x}}^2}((\nabla_{\mathbf{x}} g(\mathbf{x})^2)^T \boldsymbol{\sigma}_{\mathbf{x}}^2 - \sigma_{g(\mathbf{x})}^2)^2 \tag{5}$$

Assuming smooth $K$-neighbors [9]:

$$\text{assume } \boldsymbol{\sigma}_{\mathbf{x}_i}^2 \approx \boldsymbol{\sigma}_{\mathbf{x}_h}^2 \tag{6}$$

$$\text{thus } \boldsymbol{\sigma}_{\mathbf{x}_i}^{2*} = \arg\min_{\boldsymbol{\sigma}_{\mathbf{x}_i}^2} \sum_{h=1}^{K}((\nabla_{\mathbf{x}_h} g(\mathbf{x}_h)^2)^T \boldsymbol{\sigma}_{\mathbf{x}_i}^2 - \sigma_{g(\mathbf{x}_h)}^2)^2 \tag{7}$$

# Smoothness constrained attribution (SCA) II

## Problem solution

Define

$$\mathbf{B} = \begin{pmatrix} (\nabla_{\mathbf{x}_0} g(\mathbf{x}_0)^2)^T \\ \vdots \\ (\nabla_{\mathbf{x}_K} g(\mathbf{x}_K)^2)^T \end{pmatrix}, \mathbf{s} = \begin{pmatrix} \sigma^2_{g(\mathbf{x}_0)} \\ \vdots \\ \sigma^2_{g(\mathbf{x}_K)} \end{pmatrix},$$

and solve

$$\boldsymbol{\sigma}^2_{\mathbf{x}_i} = \underset{\boldsymbol{\sigma}^2_{\mathbf{x}_i}}{\arg\min} \left\|\mathbf{B}\boldsymbol{\sigma}^2_{\mathbf{x}_i} - \mathbf{s}\right\|_2^2 . \tag{8}$$

using the Moore-Penrose inverse [10]–[12]:

$$\boldsymbol{\sigma}^{2*}_{\mathbf{x}_i,\text{unc.}} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{s}. \tag{9}$$

# Extension

## $\ell_2$-regularization

$\ell_2$ penalty

$$\boldsymbol{\sigma}_{\mathbf{x}_i}^{2*} = \arg\min_{\boldsymbol{\sigma}_{\mathbf{x}_i}^2} \left\|\mathbf{B}\boldsymbol{\sigma}_{\mathbf{x}_i}^2 - \mathbf{s}\right\|_2^2 + \frac{\lambda}{2}\left\|\boldsymbol{\sigma}_{\mathbf{x}_i}^2\right\|_2^2 \tag{10}$$

yields the regularized solution

$$\mathbf{a}(\mathbf{x}_i) = \boldsymbol{\sigma}_{\mathbf{x}_i,\ell_2}^{*} = \sqrt{(\mathbf{B}^T\mathbf{B} + \lambda\mathbf{I})^{-1}\mathbf{B}^T\mathbf{s}}. \tag{11}$$

# Extension

## $\ell_2$-regularization

$\ell_2$ penalty

$$\boldsymbol{\sigma}^{2*}_{\mathbf{x}_i} = \arg\min_{\boldsymbol{\sigma}^2_{\mathbf{x}_i}} \left\| \mathbf{B}\boldsymbol{\sigma}^2_{\mathbf{x}_i} - \mathbf{s} \right\|_2^2 + \frac{\lambda}{2} \left\| \boldsymbol{\sigma}^2_{\mathbf{x}_i} \right\|_2^2 \tag{10}$$

yields the regularized solution

$$\mathbf{a}(\mathbf{x}_i) = \boldsymbol{\sigma}^*_{\mathbf{x}_i,\ell_2} = \sqrt{(\mathbf{B}^T\mathbf{B} + \lambda\mathbf{I})^{-1}\mathbf{B}^T\mathbf{s}}. \tag{11}$$

? Unknown $\nabla_{\mathbf{x}} g(\mathbf{x})$ & $\sigma^2_{g(\mathbf{x})}$

# Implementation

## Assumptions

$$\nabla_{\mathbf{x}} g(\mathbf{x}) \approx \Phi(f, \mathbf{x}). \tag{12}$$

$$\sigma^2_{g(\mathbf{x})} \approx U_a(\mathbf{x})^2. \tag{13}$$

- ✔ Eq.12: Use robust XAI [13]
- ✔ Eq.13: Use calibrated UCQ [14]

# Implementation

## Assumptions

$$\nabla_{\mathbf{x}} g(\mathbf{x}) \approx \Phi(f, \mathbf{x}). \tag{12}$$

$$\sigma^2_{g(\mathbf{x})} \approx U_a(\mathbf{x})^2. \tag{13}$$

- ✔ Eq.12: Use robust XAI [13]
- ✔ Eq.13: Use calibrated UCQ [14]

# Implementation

## Assumptions

$$\nabla_{\mathbf{x}} g(\mathbf{x}) \approx \Phi(f, \mathbf{x}). \tag{12}$$

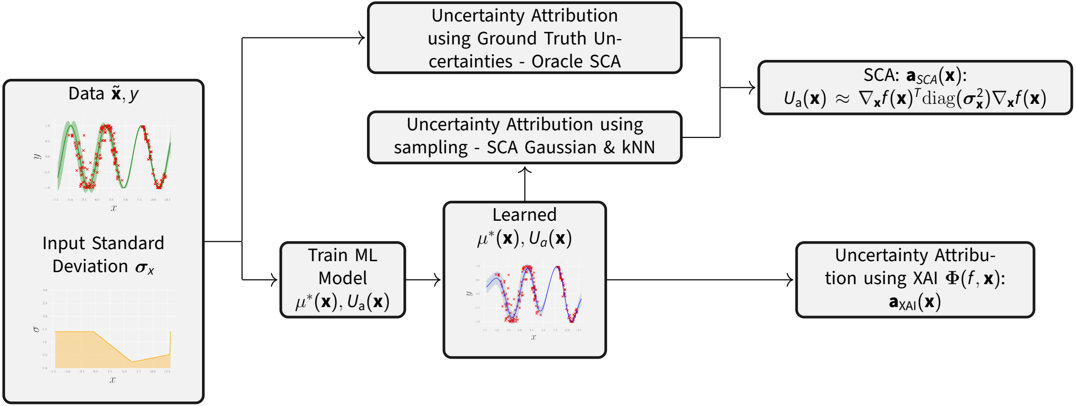
$$\sigma^2_{g(\mathbf{x})} \approx U_a(\mathbf{x})^2. \tag{13}$$

- ✔ Eq.12: Use robust XAI [13]
- ✔ Eq.13: Use calibrated UCQ [14]

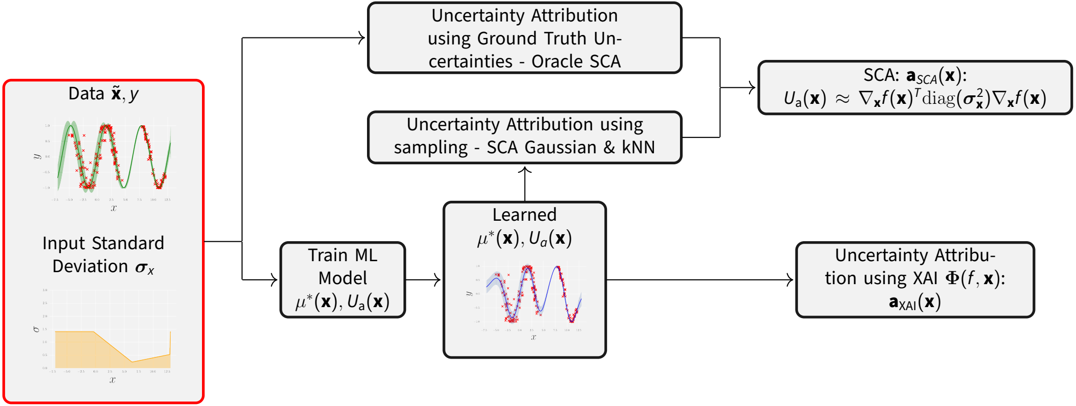
# Implemented input uncertainty attribution mechanisms (iUCAMs)

- Smoothness constrained attribution (SCA) Gaussian: sample $K$ Gaussian neighbors
- SCA kNN: find $k$ closest neighbors
- SCA Oracle: Use “ideal” model - ground truth $\nabla_{\mathbf{x}} g(\mathbf{x})$ & $\sigma_{g(\mathbf{x})}$
- Predict 0: simplest baseline
- Gradient-based / SHAP: explain $U_a(\mathbf{x})$ [5], [6]

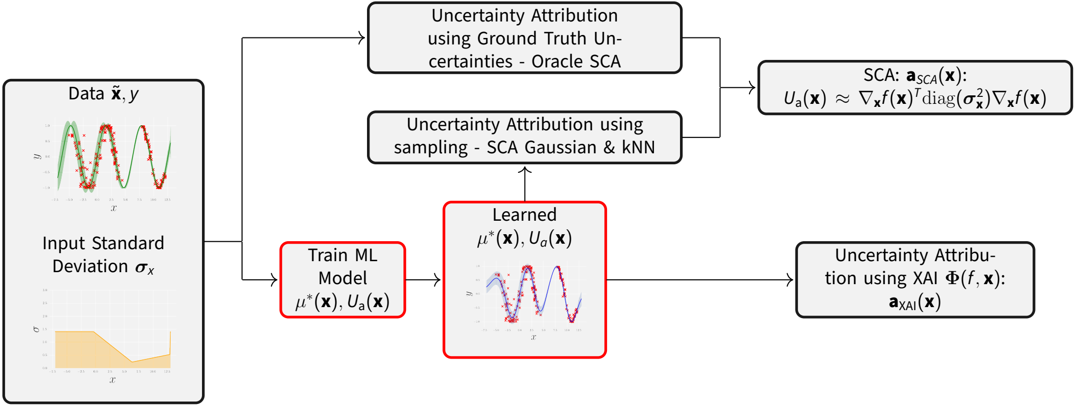
# Overview iUCAM



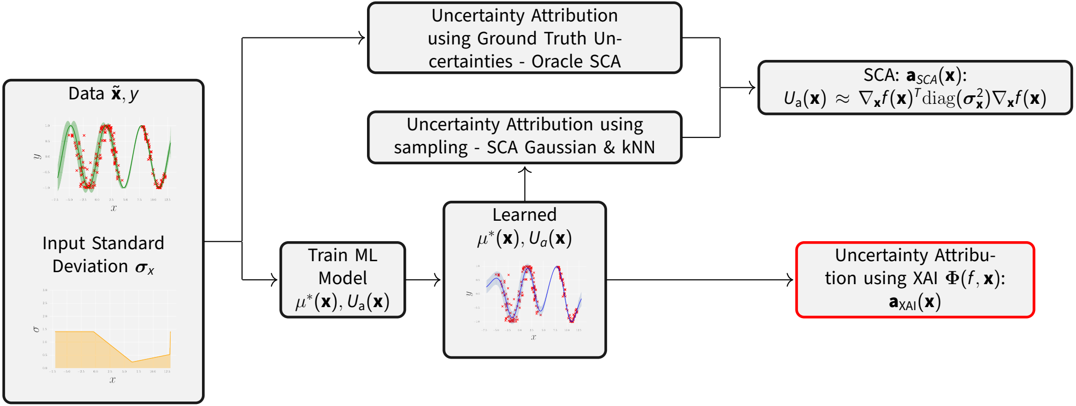
# Overview iUCAM



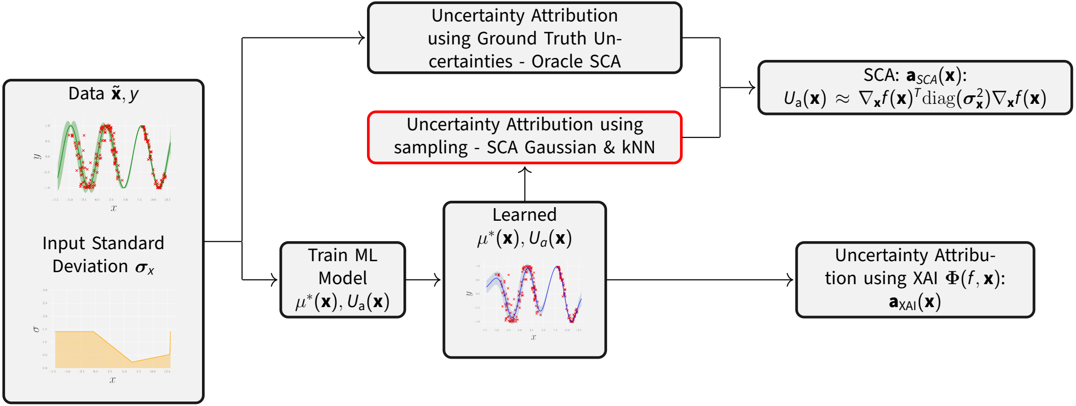
# Overview iUCAM



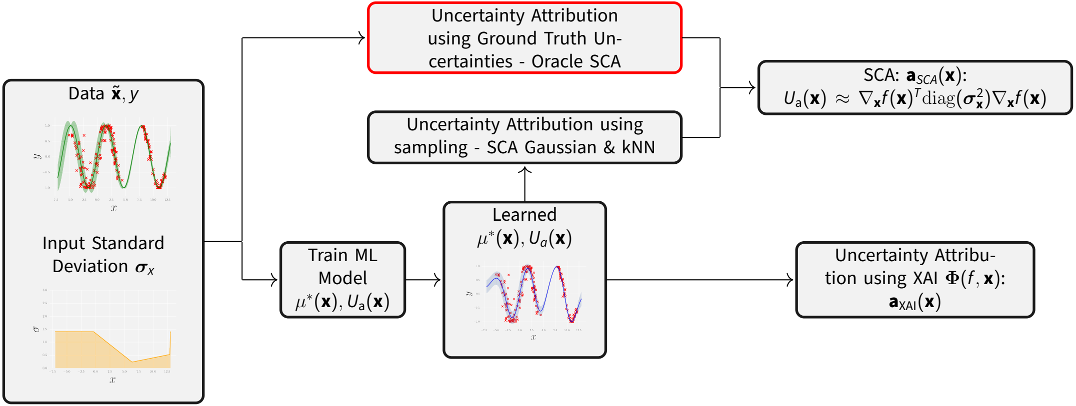
# Overview iUCAM



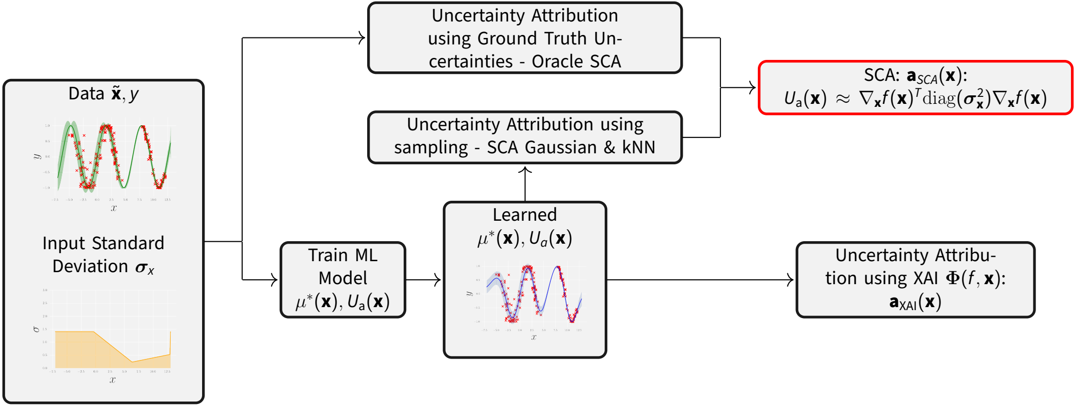
# Overview iUCAM



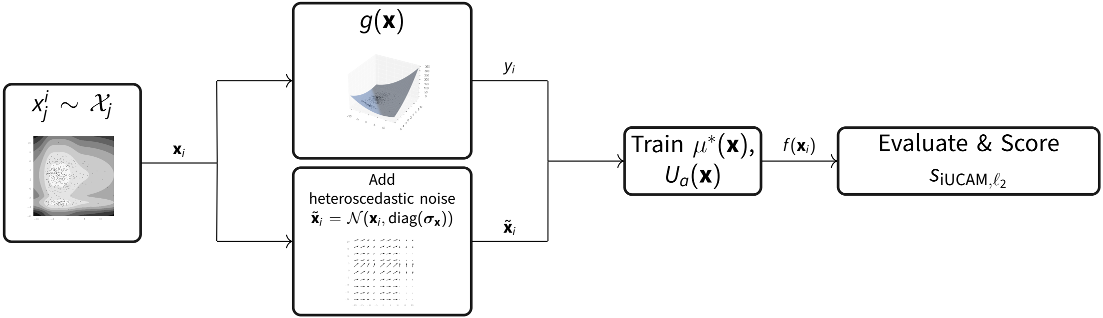
# Overview iUCAM



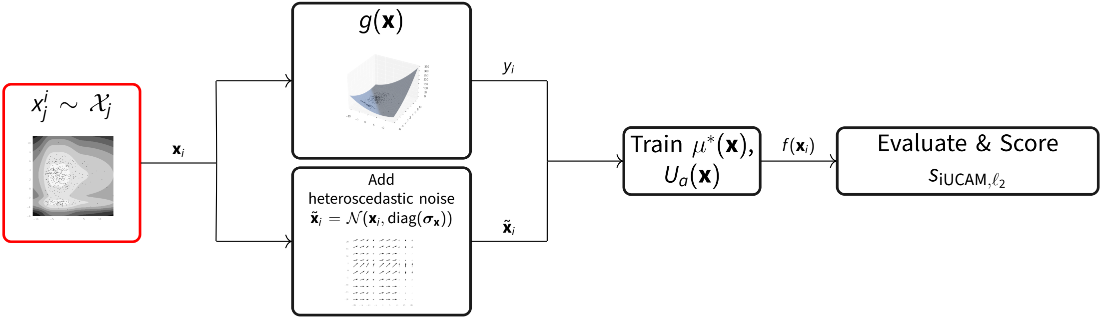
# Overview iUCAM



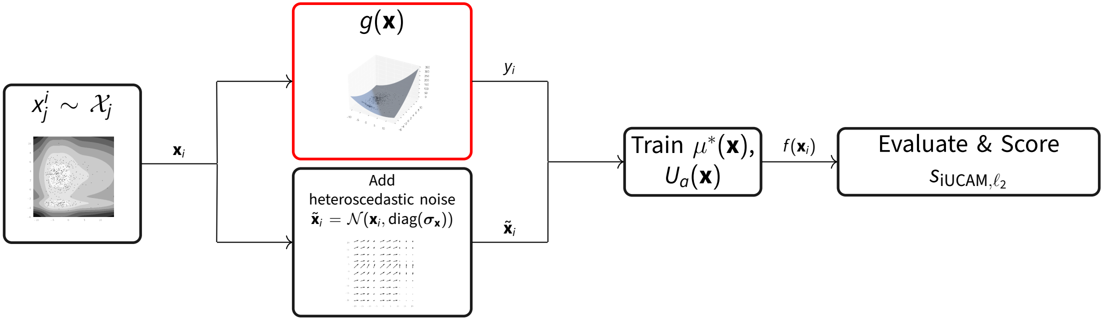
# Evaluation overview - *Inverse Crime*



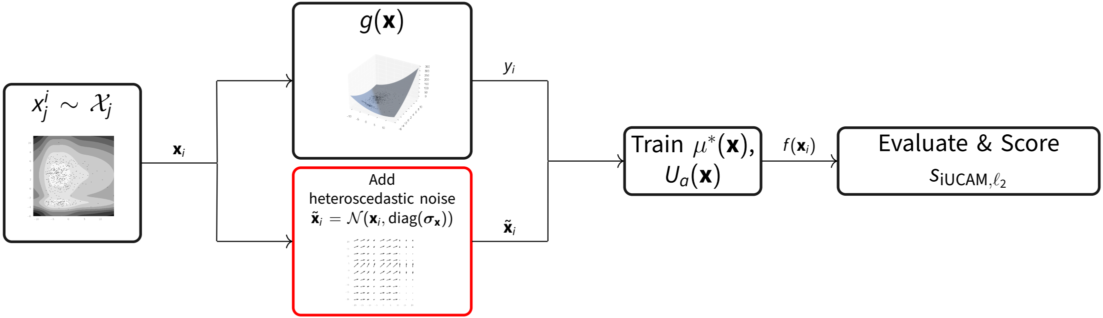
# Evaluation overview - *Inverse Crime*



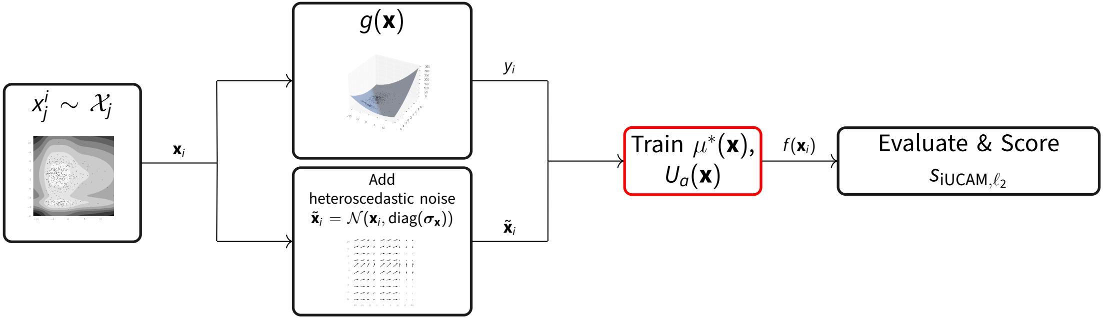
# Evaluation overview - *Inverse Crime*



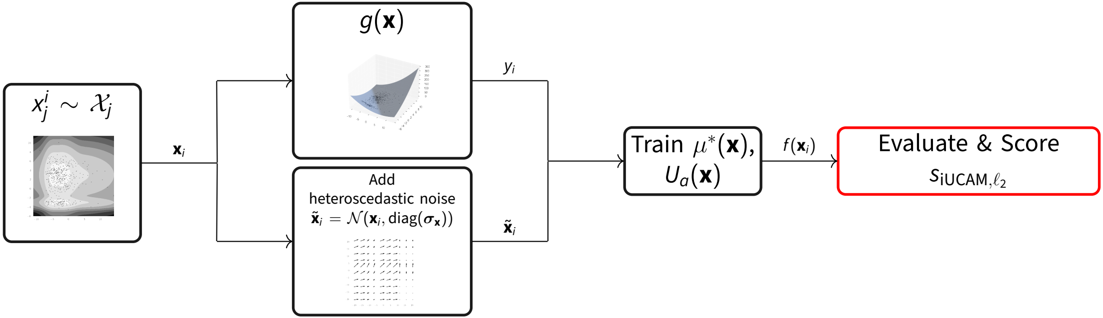
# Evaluation overview - *Inverse Crime*



# Evaluation overview - *Inverse Crime*

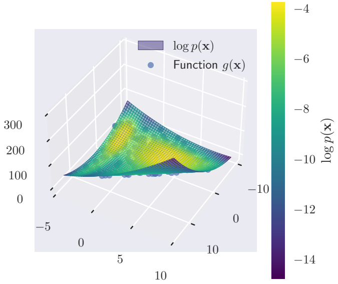


# Evaluation overview - *Inverse Crime*



# Datasets

## Toy Datasets [15]




### Toy Datasets

- Polynomial

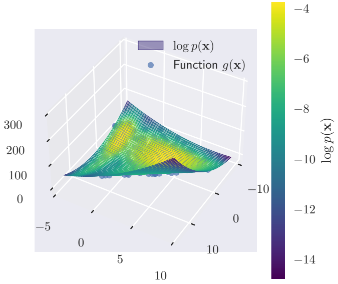
$$g(x_1, x_2) = k_1 x_1^2 + k_2 x_2^2 + k_3 x_1 x_2 + k_4 x_1 + k_5 x_2 + k_6$$

- Styblinski-Tang [15]

$$g(\mathbf{x}) = \frac{1}{2} \sum_{j=1}^{d} k_j (x_j^4 - 16 x_j^2 + 5 x_j)$$

# Datasets

## Toy Datasets [15]




### Toy Datasets

- Polynomial

$$g(x_1, x_2) = k_1 x_1^2 + k_2 x_2^2 + k_3 x_1 x_2 + k_4 x_1 + k_5 x_2 + k_6$$

- Styblinski-Tang [15]

$$g(\mathbf{x}) = \frac{1}{2} \sum_{j=1}^{d} k_j (x_j^4 - 16 x_j^2 + 5 x_j)$$

# iUCAM Visualization

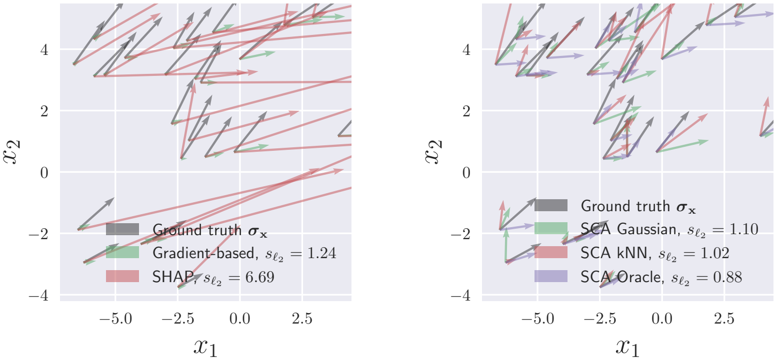
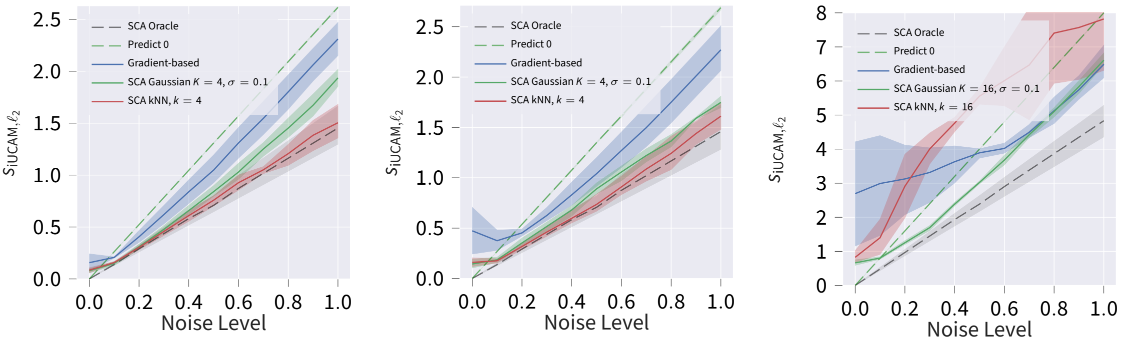



Figure: Visualized vectors of the iUCAM, for $l = 0.60$ on the polynomial Gaussian mixture model (GMM) with the light gradient boosting machine (LightGBM).

# iUCAM Results




(a) Polynomial (b) Polynomial GMM (c) Styblinski-Tang, $d = 4$

Figure: Results of the uncertainty attribution noise sweep using the LightGBM.

# Conclusion

- Smoothness constrained attribution (SCA): model-agnostic & source free iUCAM

# Conclusion

- Smoothness constrained attribution (SCA): model-agnostic & source free iUCAM
- SCA performs well on lower dimensional datasets

# Conclusion

- Smoothness constrained attribution (SCA): model-agnostic & source free iUCAM
- SCA performs well on lower dimensional datasets
- Future work: improve hyperparameters for high dimensional data

TU Graz

# Bibliography I


[1] F. Doshi-Velez and B. Kim, **Towards a rigorous science of interpretable machine learning,** 2017. arXiv: 1702.08608 [stat.ML].

[2] U. Bhatt, J. Antorán, Y. Zhang, *et al.*, **Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty,** Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, ser. AIES '21, Virtual Event, USA: Association for Computing Machinery, 2021, pp. 401–413, ISBN: 9781450384735. DOI: 10.1145/3461702.3462571. [Online]. Available: https://doi.org/10.1145/3461702.3462571.

[3] L. R. Marusich, J. Z. Bakdash, Y. Zhou, and M. Kantarcioglu, **Using ai uncertainty quantification to improve human decision-making,** 2024. arXiv: 2309.10852 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2309.10852.

# Bibliography II


[4] J. C. Cresswell, Y. Sui, B. Kumar, and N. Vouitsis, **Conformal prediction sets improve human decision making,** 2024. arXiv: 2401.13744 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2401.13744.

[5] H. Wang, D. Joshi, S. Wang, and Q. Ji, **Gradient-based uncertainty attribution for explainable bayesian deep learning,** Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12 044–12 053.

[6] P. Iversen, S. Witzke, K. Baum, and B. Y. Renard, **Identifying drivers of predictive aleatoric uncertainty,** 2024. arXiv: 2312.07252.

[7] J. Antoran, U. Bhatt, T. Adel, A. Weller, and J. M. Hernández-Lobato, **Getting a {clue}: A method for explaining uncertainty estimates,** International Conference on Learning Representations, 2021. [Online]. Available: https://openreview.net/forum?id=XSLF1XFq5h.

# Bibliography III


[8] K. O. Arras, **An introduction to error propagation: Derivation, meaning and examples of equation cy= fx cx fxt,** ETH Zurich, Tech. Rep., 1998.

[9] B. D. Lucas and T. Kanade, **An iterative image registration technique with an application to stereo vision,** Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2, ser. IJCAI'81, Vancouver, BC, Canada: Morgan Kaufmann Publishers Inc., 1981, pp. 674–679.

[10] A. Dresden, **The fourteenth western meeting of the American Mathematical Society,** *Bulletin of the American Mathematical Society*, vol. 26, no. 9, pp. 385–396, 1920.

[11] A. Bjerhammar, **Application of calculus of matrices to method of least squares : With special reference to geodetic calculations,**, 1951. [Online]. Available: https://api.semanticscholar.org/CorpusID:118134976.

TU Graz

# Bibliography IV


[12] R. Penrose, **A generalized inverse for matrices,** *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 51, no. 3, pp. 406–413, 1955. DOI: 10.1017/S0305004100030401.

[13] B. Kantz, C. Staudinger, C. Feilmayr, *et al.*, **Robustness of explainable artificial intelligence in industrial process modelling,** 2024. arXiv: 2407.09127 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2407.09127.

[14] A. N. Angelopoulos and S. Bates, **A gentle introduction to conformal prediction and distribution-free uncertainty quantification,** 2022. arXiv: 2107.07511 [cs.LG].

[15] M. Styblinski and T.-S. Tang, **Experiments in nonconvex optimization: Stochastic approximation with function smoothing and simulated annealing,** *Neural Networks*, vol. 3, no. 4, pp. 467–483, 1990, ISSN: 0893-6080. DOI: https://doi.org/10.1016/0893-6080(90)90029-K. [Online]. Available: https://www.sciencedirect.com/science/article/pii/089360809090029K.

# Causal Approaches

- ? Understand process
- → Causal learning?
- ? Unknown ground truths, bad performance

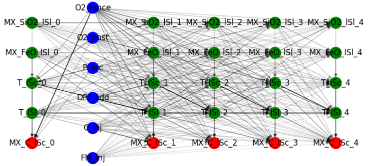


Figure: “Ground truth” Causal Graph

# Regression scoring

## $s_{R^2}$

$$s_{R^2} = 1 - \frac{\sum_{i=1}^{n}(\hat{\mu}(\mathbf{x}_i) - y_i)^2}{\sum_{i=1}^{n}(\mu - y_i)^2}$$

$$\begin{aligned} \mathbf{x}_i &\ldots \text{input data sample} \\ \hat{\mu}(\mathbf{x}_i) &\ldots \text{prediction} \\ y_i &\ldots \text{Target:} \\ \mu &\ldots \text{mean of Target: s} \sum_{i=1}^{n} y_i \end{aligned}$$

Target: $s_{\text{XAI}} \approx 1$, worse than mean predictor if $s_{\text{XAI}} < 0$

# iUCAM scoring

## $s_{\text{iUCAM},\ell_2}$

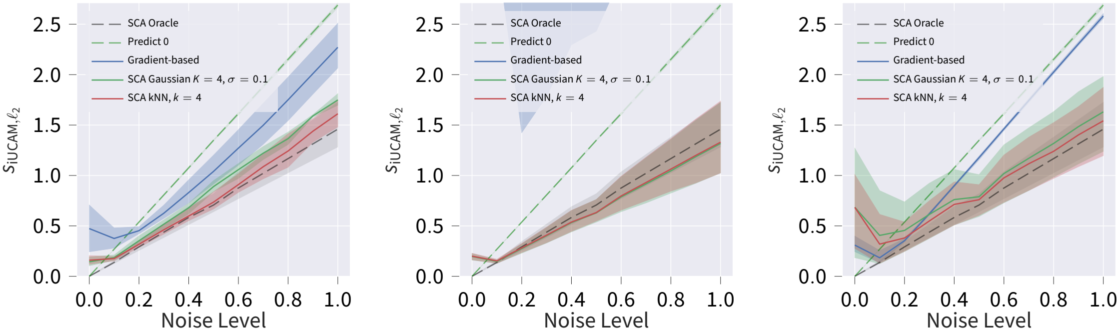
$$s_{\text{iUCAM},\ell_2} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{a}(\mathbf{x}_i) - \boldsymbol{\sigma}_{\mathbf{x}_i}\|$$

$\mathbf{a}(\mathbf{x}_i)$ . . . attribution

$\boldsymbol{\sigma}_{\mathbf{x}_i}$ . . . ground truth heteroscedastic noise

Target: $s_{\text{iUCAM},\ell_2} \approx 0$

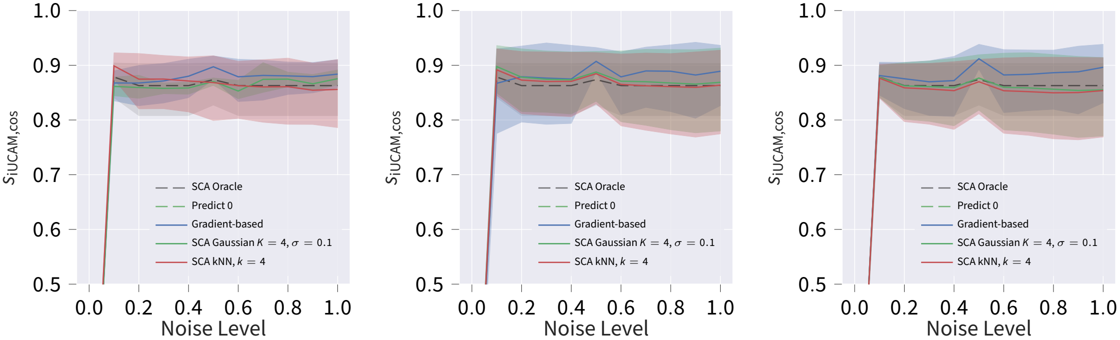
# More iUCAM Results I




(a) LightGBM (b) neural network (NN) (c) Gaussian process (GP)

**Figure:** Results of the uncertainty attribution noise sweep using on the Polynomial GMM - $s_{\text{iUCAM},\ell_2}$.

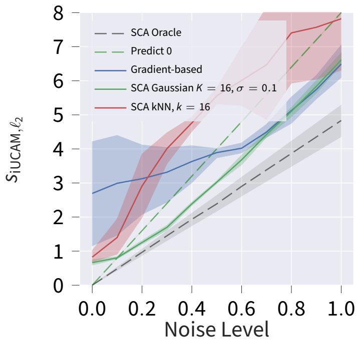
# More iUCAM Results II



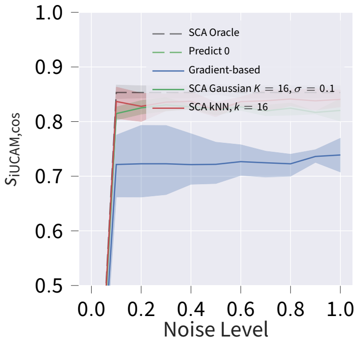

(a) LightGBM (b) NN (c) GP

Figure: Results of the uncertainty attribution noise sweep using on the Polynomial GMM - $s_{\text{iUCAM,cos}}$.

# More iUCAM Results III




(a) $s_{\mathrm{iUCAM},\ell_2}$

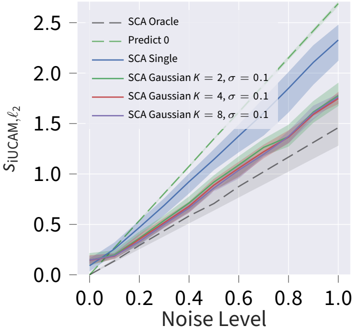



(b) $s_{\mathrm{iUCAM,cos}}$

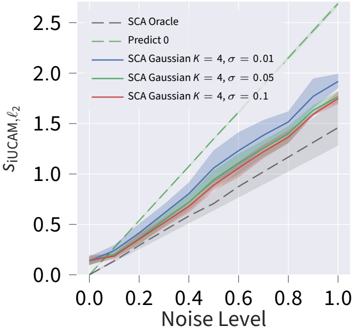
Figure: Results of the uncertainty attribution noise sweep using on the Styblinski-Tang, $d = 4$ dataset & LightGBM.

# More iUCAM Results IV

# SCA Ablation Studies I



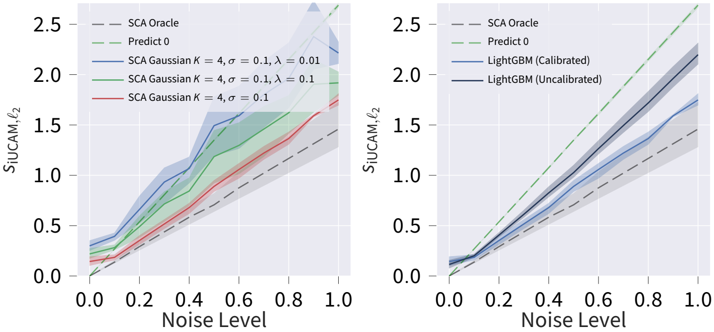

(a) Varying $K$




(b) Varying $\sigma$

Figure: Ablation studies on the SCA with the Polynomial GMM.

# SCA Ablation Studies II




(a) Varying $\lambda$ (b) Calibration

Figure: Ablation studies on the SCA with the Polynomial GMM.