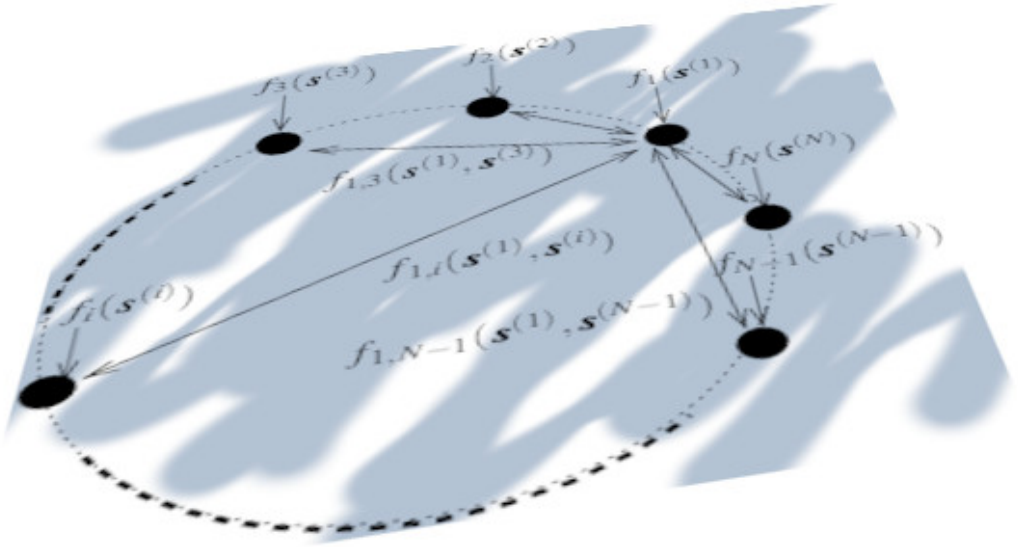
Analyzing Complex Models by Orthogonal Input-Output Decompositions



Pavel Loskot
pavelloskot@intl.zju.edu.cn

ZJU-UIUC INSTITUTE
Zhejiang University-University of Illinois at Urbana-Champaign Institute
浙江大学伊利诺伊大学厄巴纳香槟校区联合学院

The First International Conference on Systems Explainability
EXPLAINABILITY 2024
November 17, 2024 to November 21, 2024 - Valencia, Spain

About Me

Pavel Loskot joined the ZJU-UIUC Institute as Associate Professor in January 2021. He received his PhD degree in Wireless Communications from the University of Alberta in Canada, and the MSc and BSc degrees in Radioelectronics and Biomedical Electronics, respectively, from the Czech Technical University of Prague. He is the Senior Member of the IEEE, Fellow of the HEA in the UK, and the Recognized Research Supervisor of the UKCGE.

In the past 25 years, he was involved in numerous industrial and academic collaborative projects in the Czech Republic, Finland, Canada, the UK, Turkey, and China. These projects concerned mainly wireless and optical telecommunication networks, but also genetic regulatory circuits, air transport services, and renewable energy systems. This experience allowed him to truly understand the interdisciplinary workings, and crossing the disciplines boundaries.

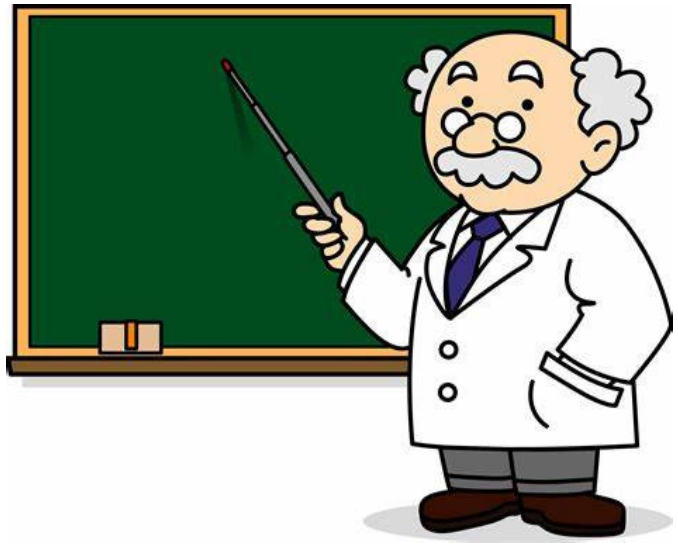
His current research focuses mathematical and probabilistic modeling, statistical signal processing and classical machine learning for multi-sensor data in biomedicine, computational molecular biology, and wireless communications.

OBJECTIVES

- explore strategies for extracting information about complex models to achieve their explainability
- consider specifically the methods for factorizations, decompositions, and expansions of univariate and multivariate functions
- investigate Sobol's decomposition of multivariate functions, which did not receive sufficient attention it likely deserves

OUTLINE

- Mathematical modeling
- Decompositions of multivariate functions
- Orthogonalizing inputs and outputs
- Numerical example
- Observations and future work



Mathematical Modeling

Modeling economics

- fast and cheap explorations
  → replacing lab/field experiments
- used everywhere nowadays
- improving information gains

Model formats

- mathematical expressions
  → reproducability
- algorithms, computer simulations
- datasets

Model interpretability

- understand how outputs are obtained from inputs
  → understand the model structure
- study how parameters affect model properties
  → sensitivity analysis
- various model-related tasks:
  → calibration, optimization, selection, validation, simplification

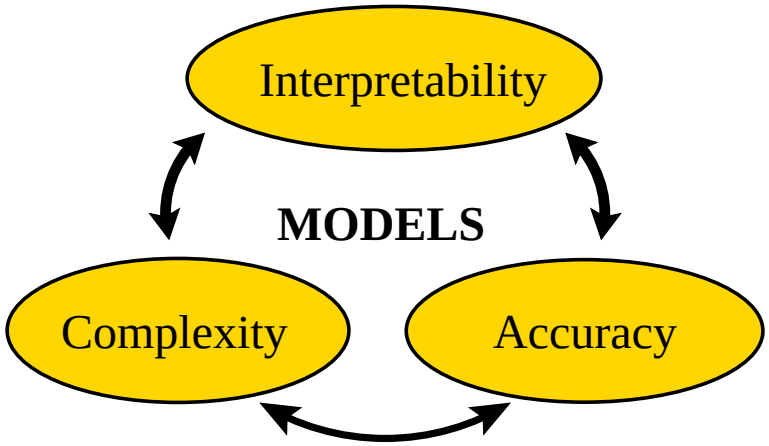
Interpretability strategies

- local and global sensitivity
- surrogate and meta models
- output (variance) expansions

Mathematical Modeling (cont.)

Model explainability

- narrower objective than interpretability
- which inputs more important for outputs
  → factor screening, attribution problem
- model-agnostic methods
  → permutation importance
  → dependency plots
  → SHAP explanations



Orthogonal decomposition

$$M(x) = a_1 M_1(x) + a_2 M_2(x) + \cdots + M_N(x)$$

- generalized Fourier series
- if components $M_i(x)$ are mutually orthogonal
  → straightforward to find coefficients a_i
  → analysis of $M_i(x)$ can be separated
  → faster learning (convergence)
  → clear high-level structure i.e. explainability
- methods
  → Gram-Schmidt process, matrix factorizations (SVD, PCA, QR, ...)

Decompositions of Multivariate Functions

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(x_1, \ldots, x_I) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_O(\boldsymbol{x}) \end{bmatrix} \in \mathcal{R}^O, \ \boldsymbol{x} \in \mathcal{R}^I$$

Objectives

- uncover latent structure
- reducing computational complexity
  → divide & conquer
- provide explainability
- approximations
  → optimization and analysis

Function factorization

$$\boldsymbol{f}(\boldsymbol{x}) = \prod_{i=1}^{n} \boldsymbol{f}_i(\boldsymbol{s}_i), \ \boldsymbol{s}_i \subseteq \{x_1, \ldots, x_I\} \quad \text{(product factors)}$$

$$\boldsymbol{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \boldsymbol{f}_i(\boldsymbol{s}_i), \ \boldsymbol{s}_i \subseteq \{x_1, \ldots, x_I\} \quad \text{(sum factors)}$$

$$\boldsymbol{f}(\boldsymbol{x})A(\boldsymbol{x}) = \begin{cases} \boldsymbol{f}(\boldsymbol{x}), & \boldsymbol{x} \in \mathcal{A} \\ 0, & \boldsymbol{x} \notin \mathcal{A} \end{cases} \quad \text{(adding constraints)}$$

Decompositions of Multivariate Functions (cont.)

Stochastic function

$$\boldsymbol{y} \approx \sum_{i=1}^{n} a_i |\boldsymbol{x} - \mathrm{E}[\boldsymbol{x}]|_1^i \quad \text{[Loskot, 2021]}$$

$\rightarrow$ multivariate Taylor expansion

Universal approximation theorem

$$\boldsymbol{y} \approx = \cdots \boldsymbol{\sigma} \circ (\boldsymbol{A}_i, \boldsymbol{b}_i) \circ \cdots \boldsymbol{\sigma}(\boldsymbol{A}_1 \boldsymbol{x} + \boldsymbol{b}_1) \quad \text{[Hornik, 1989]}$$

Kolmogorov-Arnold approximation

$$f(\boldsymbol{x}) = \sum_{i=0}^{2n} \Phi\left(\sum_{j=1}^{n} \phi_{i,j}(x_j) \right) \quad \text{[Lorentz, 1962]}$$

Sobol's decomposition

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}_0 + \sum_{i=1}^{I} \boldsymbol{f}_i(\boldsymbol{x}_i) + \sum_{\substack{i,j=1 \\ i \neq j}}^{I} \boldsymbol{f}_{i,j}(\boldsymbol{x}_i, \boldsymbol{x}_j) \cdots + \sum_{i=1}^{I} \boldsymbol{f}_{\{1:I\} \setminus i}(\boldsymbol{x})$$

Orthogonal Input-Output Decompositions

Main task

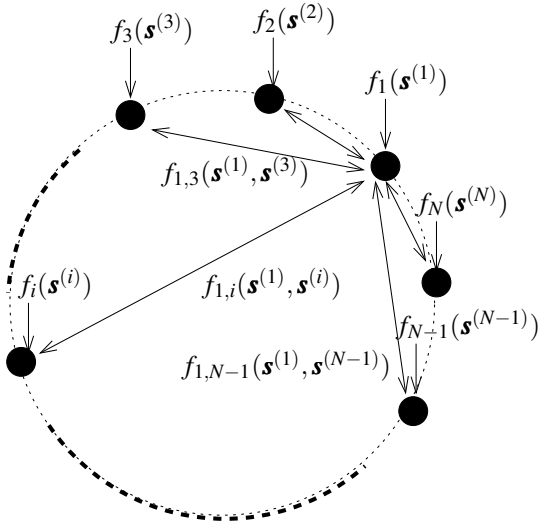
- analyze explainability of complex model, $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\Omega}) \equiv M(\boldsymbol{x};\boldsymbol{\Omega}) \equiv M_{\boldsymbol{\Omega}}(\boldsymbol{x})$

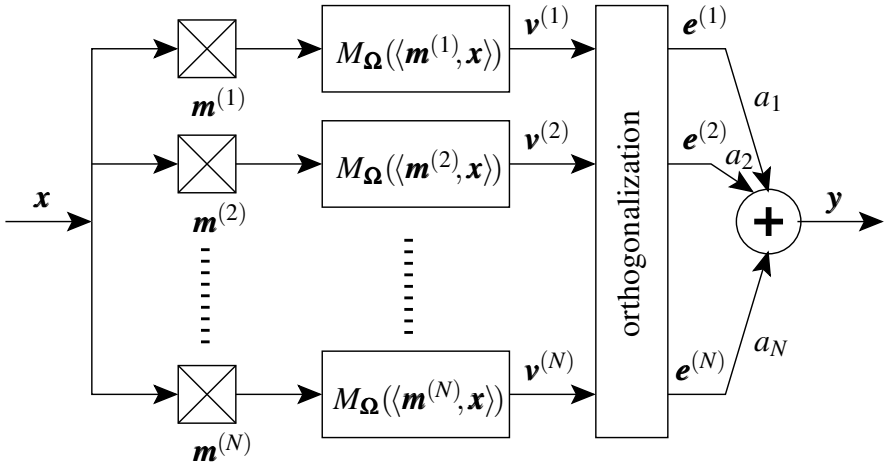
Key paper contribution

$$\boldsymbol{f}(\boldsymbol{x}) \approx \boldsymbol{f}_0 + \sum_{i=1}^{N} \boldsymbol{f}_i(\boldsymbol{s}^{(i)}) + \sum_{\substack{i,j=1 \\ i\neq j}}^{N} \boldsymbol{f}_{i,j}(\boldsymbol{s}^{(i)}, \boldsymbol{s}^{(j)})$$

orthogonal input projections (masking)

$$\begin{cases} \boldsymbol{s}^{(i)} \cap \boldsymbol{s}^{(j\neq i)} = \emptyset & \text{(disjoint)} \\ \cup_i \boldsymbol{s}^{(i)} = \boldsymbol{x} & \text{(full coverage)} \end{cases}$$





Numerical Example

Problem

- classifying hand-written digits in MNIST dataset
- train a MLP with two hidden layers
  $\rightarrow$ training accuracy 97.56%, testing accuracy 94.31%

Explore

1. training dataset vs. randomly generated inputs
2. orthogonal (disjoint) masking vs. random masking of inputs

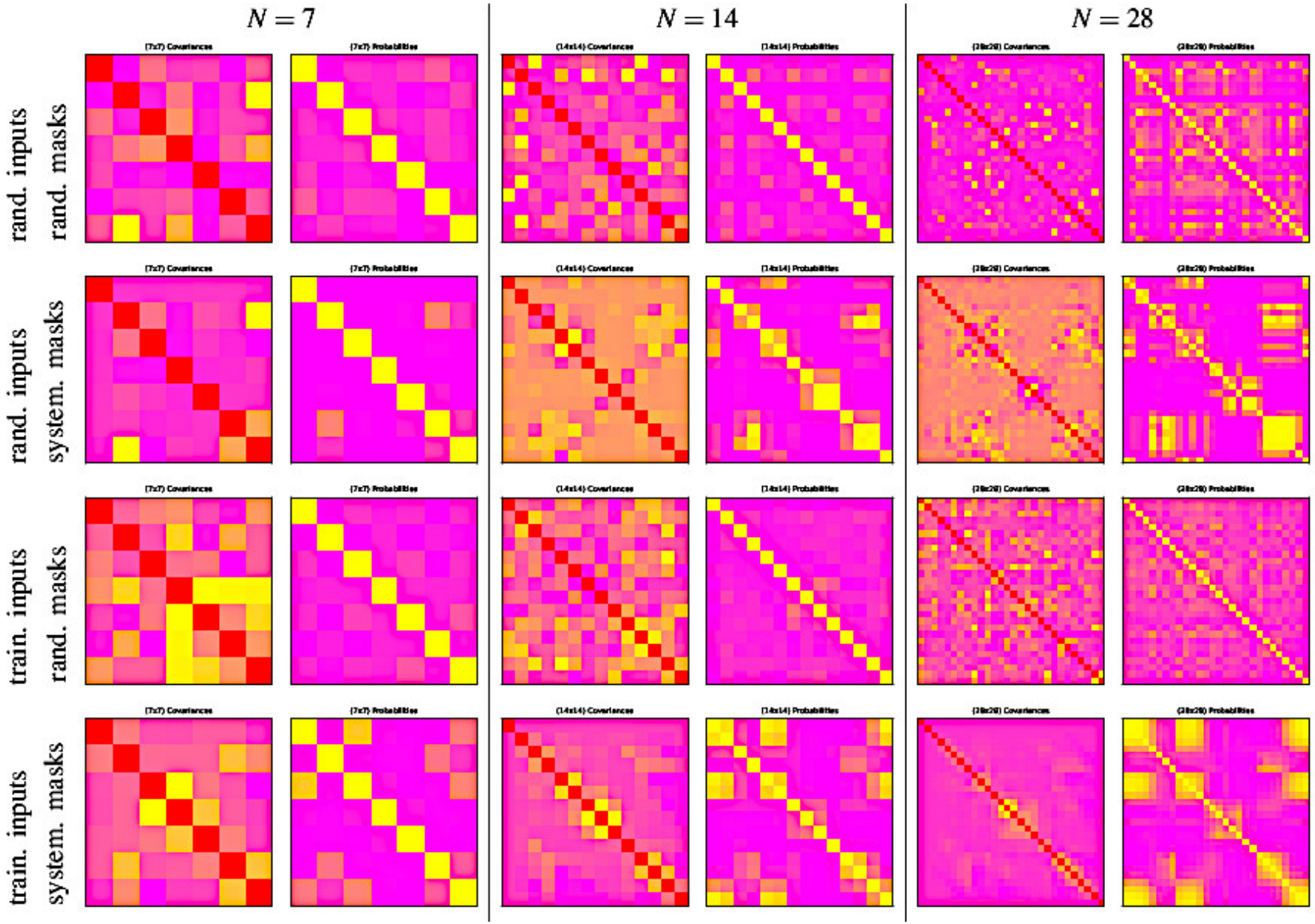
$$\overline{\mathrm{MSE}}_0 = \frac{1}{K}\sum_{k=1}^{K} \|\boldsymbol{y}(k)\|^2$$

$$\overline{\mathrm{MSE}}_1 = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K} \left\|\boldsymbol{v}^{(i)}(k) - \boldsymbol{y}(k)\right\|^2$$

$$\min \mathrm{MSE}_1 = \min_{1\le i\le N} \frac{1}{K}\sum_{k=1}^{K} \left\|\boldsymbol{y}^{(i)}(k) - \boldsymbol{y}(k)\right\|^2$$

$$\overline{\mathrm{MSE}}_2 = \frac{1}{K}\sum_{k=1}^{K} \|\tilde{\boldsymbol{y}}(k) - \boldsymbol{y}(k)\|^2$$

Numerical Example (cont.)



→ correlations and probabilities (decisions with masks $\boldsymbol{m}^{(i)}$ and $\boldsymbol{m}^{(j)}$ the same as with combined mask $\boldsymbol{m}^{(i)} + \boldsymbol{m}^{(j)}$

Observations and Conclusion

MSE values

- input masking substantially reduces accuracy in exchange for interpretability
- combining outputs corresponding to masks inputs restores some accuracy
- larger sensitivity for training vs. random inputs than for orthogonal vs. random masks

Correlations vs. probabilities

- increasing N increases resolution and variance of calculated values
- some orthogonal inputs have dependent decisions (square patterns)
- probability patterns require sufficiently large N to be visible
  $\rightarrow$ i.e. explainability requires the minimum resolution

Summary

- Sobol-based decomposition of multivariate functions
  $\rightarrow$ graph representation of multivariate functions
- input masking
  $\rightarrow$ projections into orthogonal input subspaces
- output orthogonalization via decorrelation
- SVD of correlation matrix
- easy to obtain linear combining coefficients

Future Work

Some ideas

- explainability methods must be much simpler than the model investigated
- optimizing input projections (granularity vs. complexity vs. accuracy)
  → different explainability objectives
- averaging out some inputs instead of using default values (zero)
- other strategies for obtaining component models from the base model
- constructing complex models from orthogonal component models
  → machine learning architectures
- model decomposition as structural causal model
  → causal inferences
- Sobol's decomposition
  → assuming higher-order graphs
  → approximation guarantees

The End

Thank you!

pavelloskot@intl.zju.edu.cn