Additive Gaussian Processes Revisited

Xiaoyu Lu, Alexis Boukouvalas, James Hensman

Amazon

Motivation

Consider a problem where

- Input $\mathbf{x} = (x_1, x_2, \cdots, x_D)$
- Output y

Goal: build an explainable model $y = f(\mathbf{x})$ that can be explained by each input feature x_i.

Motivation

Question: What's the form of the decomposition of f?

- $f(\mathbf{x}) = \sum_{i=1}^{d} f_i(x_i)$?
- $f(\mathbf{x}) = f_1(x_1) + f_3(x_3) + f_{12}(x_1, x_2)$?
- $f(\mathbf{x}) = f_{123}(x_1, x_2, x_3)$?

Propose Gaussian Process (GP) based model:

- good prediction performance
- lower order terms
- interpretable decomposition/visualisation

Gaussian Process Models

Definition: A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{1}$$

Additive structure of the functions and kernels:
The additive structure of the function decomposition is enforced through the structure of the kernel:

$$\begin{aligned} f(\mathbf{x}) &= f_1(x_1) + f_3(x_3) + f_{12}(x_1, x_2) \\ \iff K(x, x') &= K_1(x_1, x_1') + K_3(x_3, x_3') + K_{12}([x_1, x_2], [x_1', x_2']) \end{aligned}$$

Goal

Consider input $\mathbf{x} = (x_1, x_2, \cdots, x_D)$ and output y, we aim to build interpretable additive *Gaussian Process* (GP) model f of the form

$$y = f(\mathbf{x}) + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and

$$f(\mathbf{x}) = f_1(x_1) + f_2(x_2) + \cdots + f_{12}(x_1, x_2) + \cdots + f_{12\ldots D}(x_1, x_2, \cdots x_D)$$

Is High-Dimensional Representation Really Necessary?

On an **8-dimensional** regression problem (*Pumadyn*), *GP* with *Orthogonal Additive Kernel* (OAK) only requires

- **two 1-dimensional** main effect and
- **one 2-dimensional** interaction effect

for competitive performance.

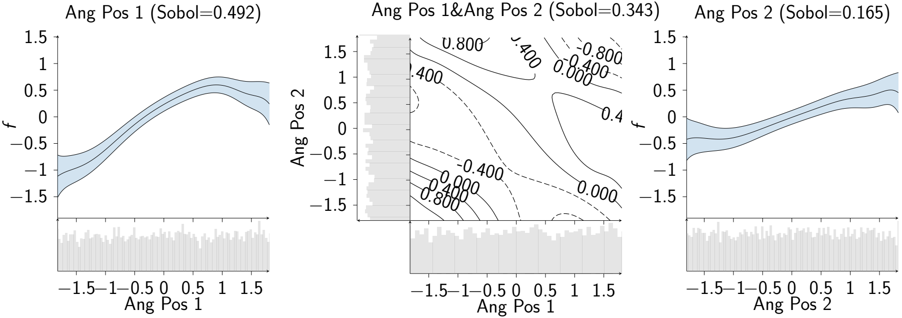


Figure: Visualization of the decomposed functions with highest Sobol indices for the pumadyn dataset. Over 99% of the variance can be explained with only these three terms.

Why it Appears to be High-Dimensional? — Orthogonality

Problem: If

$$f(x_1, x_2) = f_1(x_1) + f_{12}(x_1, x_2), \tag{2}$$

then $f_1 + \delta$, $f_{12} - \delta$ are correct decompositions for any value of δ (Märtens, 2019).

Why it Appears to be High-Dimensional? — Orthogonality

$$f(x_1, x_2) = x_1^2 - 2x_2 + \cos(3x_1)\sin(5x_2) \qquad (3)$$

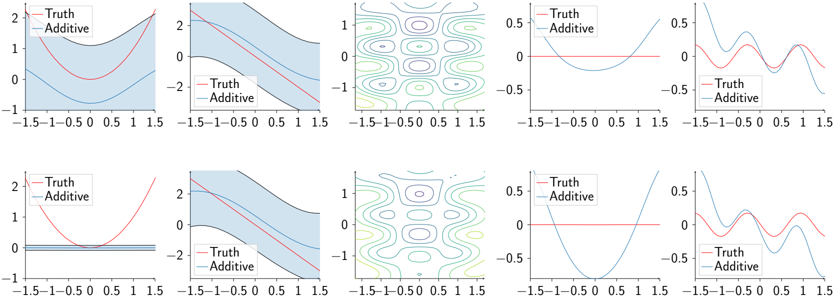


Figure: Illustration of the non-identifiability of the additive GP model in Duvenaud et al. (2011) on the two-dimensional problem, for two different decomposition with the same predictive performance.

How to Circumvent it?

We can get low-dimensional representation of

$$y = f_1(x_1) + f_2(x_2) + \cdots + f_{12}(x_1, x_2) + \cdots + f_{12\ldots D}(x_1, x_2, \cdots x_D)$$

with

- Orthogonality Constraints (Durrande et al., 2012)
- Scalability for Additive Models (Duvenaud et al., 2011)
- Sobol Index as Measure of Importance (Owen, 2014)

Constrained Kernel

Denote the density of input x_i with $p(x_i)$ and kernel of f_i with k_i, we enforce orthogonality constraints:

$$\int f_i(x_i)p(x_i)dx_i = 0 \quad \forall i, \tag{4}$$

$$\int f_{ij}(x_i, x_j)p(x_i)dx_i = 0 \quad \forall i, j \tag{5}$$

$$\cdots \tag{6}$$

$$\implies f_i \sim \mathcal{GP}(0, \tilde{k}_i).$$

If

- $p(x_i)$ is uniform, (mixture) of Gaussian or approximated using empirical dsitribution
- base kernel k_i is squared exponential kernel for continuous feature or coregional kernel for categorical feature

then $\tilde{k}_i$ is **analytic** and can be easily plugged in popular GP code.

Orthogonal Additive Kernel (OAK)

$$y = f_1(x_1) + f_2(x_2) + \cdots + f_{12}(x_1, x_2) + \cdots + f_{12\ldots D}(x_1, x_2, \cdots x_D)$$

where

$$f_i(x_i) \textit{ has kernel } \sigma_1^2 \tilde{k}_i(x_i) \tag{7}$$

$$f_{ij}(x_i, x_j) \textit{ has kernel } \sigma_2^2 \tilde{k}_i(x_i)\tilde{k}_j(x_j) \tag{8}$$

$$f_{ijk}(x_i, x_j, x_k) \textit{ has kernel } \sigma_3^2 \tilde{k}_i(x_i)\tilde{k}_j(x_j)\tilde{k}_k(x_k) \tag{9}$$

$$\cdots \tag{10}$$

- Newton-Girard trick allows for polynomial time complexity $\mathcal{O}(D^2)$

Newton-Girard Algorithm

Input: input dimension D
Input: maximum interaction order $\tilde{D}$
Input: base kernels $k_d(\cdot, \cdot), d = 1 \ldots D$
Input: order variances $\sigma_l, l = 0 \ldots \tilde{D}$
Data: input data $\mathbf{X}$
Output: kernel matrix $\mathbf{K}$
for $d = 1 \ldots D$ **do**
  $\mathbf{K}_d[i,j] = k_d(x_{i,d}, x_{j,d})$
end for
for $\ell = 0 \ldots \tilde{D}$ **do**
  $\mathbf{S}_\ell = \sum_{d=1}^{D} \mathbf{K}_d^\ell$
end for
$\mathbf{E}_0 = \mathbf{1}^{[N,N]}$
for ℓ=1... $\tilde{D}$ **do**
  $\mathbf{E}_\ell = \frac{1}{\ell} \sum_{k=1}^{\ell} (-1)^{k-1} \mathbf{E}_{\ell-k} \odot \mathbf{S}_k$
end for
$\mathbf{K} = \sum_{\ell=0}^{\tilde{D}} \sigma_\ell \times \mathbf{E}_\ell$

Illustration

$$f(x_1, x_2) = x_1^2 - 2x_2 + \cos(3x_1)\sin(5x_2) \tag{11}$$

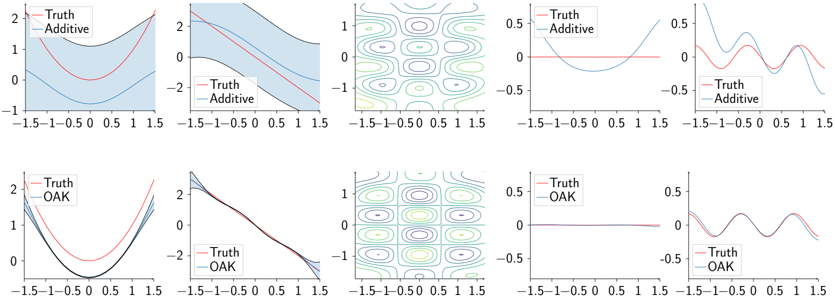


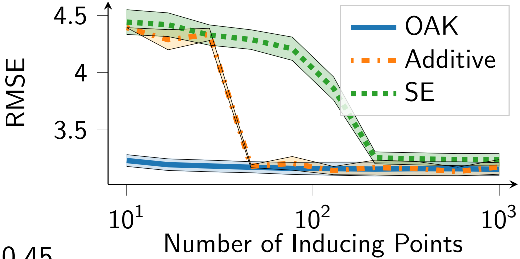
Figure: Illustration of the non-identifiability of the additive GP model in Duvenaud et al. (2011) on the two-dimensional problem. Top row: additive GP model; bottom row: OAK model.

Sparse GP with Inducing Points

- Burt et al. (2019) showed that the number of inducing points M needed is $M = \mathcal{O}(log^D N)$.
- In practice, one can limit the maximum order of interactions to be $\tilde{D} \leq D$. The number of kernels to be added for OAK is $\sum_{k=1}^{\tilde{D}} \binom{D}{k}$ and the number of inducing points needed is

$$\sum_{k=1}^{\tilde{D}} \binom{D}{k} \mathcal{O}(log^k N) = \mathcal{O}\left(\binom{D}{\tilde{D}} log^{\tilde{D}} N \right).$$

Sparse GP with Inducing Points



0.45

Figure: Test RMSE versus number of inducing points for the pumadyn dataset. Results are averaged over 5 repetitions, shaded area represents ± 1 standard deviation.

Interpretability — How to find parsimonious representation?

Q: What features (interactions) are most important?

$$f(\mathbf{x}) = f_1(x_1) + f_2(x_2) + \cdots + f_{12}(x_1, x_2) + \cdots + f_{12\ldots D}(x_1, x_2, \cdots x_D)$$

Can we decompose $\mathrm{var}_{\mathbf{x}}[f(\mathbf{x})]$?

FANOVA Decomposition and Sobol Indices

- OAK construction leads to the FANOVA decomposition:

$$f_u(\mathbf{x}) = \int_{\mathcal{X}_{-u}} \left(f(\mathbf{x}) - \sum_{v \subset u} f_v(\mathbf{x}_v) \right) dP(\mathbf{x}_{-u}), \qquad (12)$$

  where $f_\emptyset(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$, $\mathbf{x}_{-u}$ denotes $\mathbf{x}$ excluding x_u and $P(\mathbf{x})$ denotes the distribution of $\mathbf{x}$.
- The orthogonality of OAK leads to the ANOVA identity:

$$R := \mathbb{V}_{\mathbf{x}}[f(\mathbf{x})] = \sum_{u \subseteq [D]} R_u, \qquad (13)$$

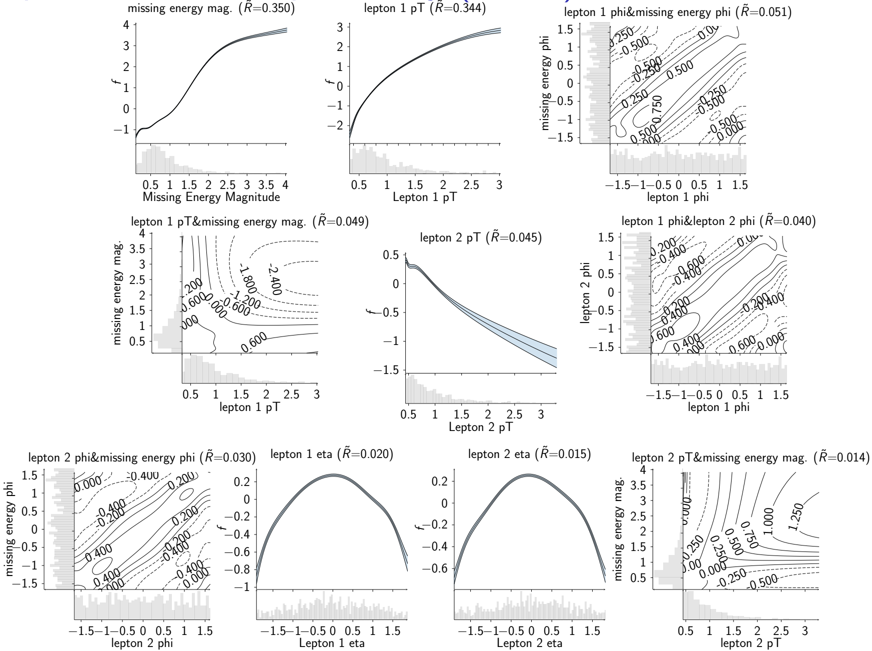
  where $R_u := \mathbb{V}_{\mathbf{x}}[f_u(\mathbf{x})]$ is defined as the Sobol index for feature set u.
- Sobol indices are **analytic** with OAK!

Normalising Flow

We use a normalizing flow to transform continuous input features to have an approximate Gaussian density:

- applying a sequence of bijective transformations on each feature
- learn the parameters of the transformation by minimizing the KL divergence between a standard Gaussian distribution and the transformed input data
- The parameters are *fixed* before fitting the OAK model on the transformed data

Experiments – Interpretability (SUSY)



Experiments – Competitive Performance

	Aggregation	OAK	Linear	SVGP	SVM	KNN	GBM	AdaBoost	MLP
Regression RMSE	avg	0.475	6.157	0.478	0.484	0.518	0.455	0.581	**0.445**
	median	0.376	0.736	0.397	0.419	0.454	**0.343**	0.580	0.361
	avg rank	3.583	6.625	4.083	4.208	4.958	**3.208**	5.750	3.583
Regression Log Likelihood	avg	**-0.229**	-0.946	-0.295	-0.585	-0.638	-0.652	-0.730	-0.891
	median	**-0.409**	-1.096	-0.512	-0.609	-0.738	-0.671	-0.875	-0.471
	avg rank	**5.583**	3.625	5.042	4.833	3.917	4.292	3.583	5.125
Classification Accuracy	avg	**0.872**	0.835	0.859	0.857	0.836	0.870	0.859	0.863
	median	0.898	0.832	0.864	0.850	0.863	**0.900**	0.892	0.873
	avg rank	**5.569**	4.224	4.741	4.500	2.983	5.224	4.207	4.552
Classification Log Likelihood	avg	**-0.267**	-0.338	-0.291	-0.306	-0.899	-0.283	-0.459	-0.306
	median	-0.280	-0.389	-0.307	-0.352	-1.088	**-0.256**	-0.584	-0.362
	avg rank	5.862	4.276	**5.931**	4.690	2.138	5.379	2.897	4.828

Figure: Average results over 24 regression datasets shown in terms of test RMSE and log likelihood (top two blocks). Average results over 29 classification datasets shown in terms of accuracy and log likelihood (bottom two blocks).

Experiments – Parsimony

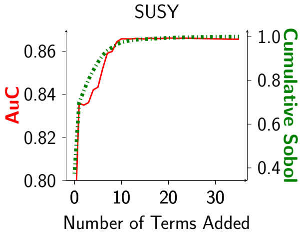


Figure: AuC as a function of number of terms added ranked by their Sobol indices for the SUSY experiments. Red solid lines and green dashed lines represent test AuC and cumulative (normalized) Sobol respectively.

Future Work

- non-independent input features
- heteroscedastic noise
- Bayesian optimisation/experimental design

Thank you!

We have open sourced our code: `https://github.com/amzn/orthogonal-additive-gaussian-processes`

References

Burt, D., Rasmussen, C. E., and Van Der Wilk, M. Rates of convergence for sparse variational Gaussian process regression. In *International Conference on Machine Learning*, pp. 862–871. PMLR, 2019.

Durrande, N., Ginsbourger, D., and Roustant, O. Additive covariance kernels for high-dimensional Gaussian process modeling. In *Annales de la faculté des sciences de Toulouse Mathématiques*, volume 21, pp. 481–499. Université Paul Sabatier, Toulouse, 2012.

Duvenaud, D. K., Nickisch, H., and Rasmussen, C. Additive Gaussian processes. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL `https://proceedings.neurips.cc/paper/2011/file/4c5bde74a8f110656874902f07378009-Paper.pdf`.

Märtens, K. *Enabling feature-level interpretability in non-linear latent variable models: a synthesis of statistical and machine learning techniques*. PhD thesis, University of Oxford, 2019.

Owen, A. B. Sobol'indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014.