UNIVERSITY OF TECHNOLOGY
IN THE EUROPEAN CAPITAL OF CULTURE
CHEMNITZ

Department of Mathematics
Applied Functional Analysis, Junior Research Group SAℓE
Franziska Nestler, Daniel Potts, Michael Schmischke, Martin Stoll and Theresa Wagner

Explainable Approximation in High Dimensions: Fourier-Based Algorithms Meet Kernel Methods

Basics

ANOVA Decomposition

Let $f \in L_2(\mathbb{T}^d)$. Based on the **ANOVA (analysis of variance) decomposition**

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{u}\subseteq\{1,\dots,d\}} f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}})$$

the importance of single dimensions as well as of groups of dimensions can be studied.

$$f_\emptyset = \int_{\mathbb{T}^d} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$

$$f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) = \int_{\mathbb{T}^{d-|\boldsymbol{u}|}} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}_{\bar{\boldsymbol{u}}} - \sum_{\boldsymbol{v}\subsetneq\boldsymbol{u}} f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}),$$

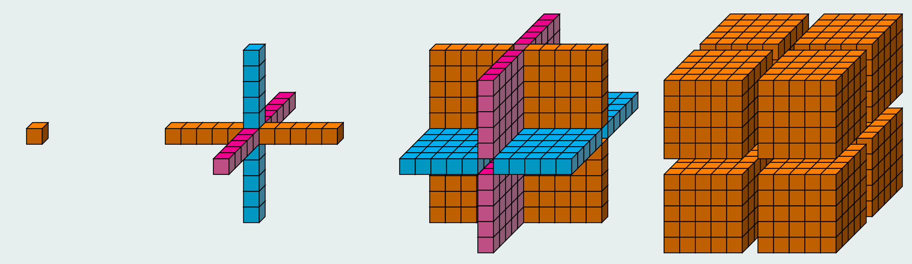
where $\bar{\boldsymbol{u}} = \{1,\dots,d\}\setminus\boldsymbol{u}$.

Decomposition in Fourier Domain

For a subset $\boldsymbol{u} \subset \{1,\dots,d\} =: \mathcal{D}$ of dimensions we have [Potts, Schmischke 2021]

$$f_{\boldsymbol{u}}(\boldsymbol{x}) = f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d:\,\mathrm{supp}(\boldsymbol{k})=\boldsymbol{u}} c_{\boldsymbol{k}}(f)\,\mathrm{e}^{2\pi\mathrm{i}\boldsymbol{k}^\top\boldsymbol{x}_{\boldsymbol{u}}}$$

Decomposition of the index set, 3D-visualization:



Variances and GSI

Consider a trigonometric polynomial

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathcal{K}} \hat{f}_{\boldsymbol{k}}\,\mathrm{e}^{2\pi\mathrm{i}\boldsymbol{k}^\top\boldsymbol{x}_j}.$$

We easily see that $\sigma^2(f) = \sum_{\emptyset\neq\boldsymbol{u}\subseteq\mathcal{D}} \sigma^2(f_{\boldsymbol{u}})$.

Study the importance of subsets $\boldsymbol{u}$ in terms of the global sensitivity indices (GSI)

$$\rho_{\boldsymbol{u}}(f) := \frac{\sigma^2(f_{\boldsymbol{u}})}{\sigma^2(f)} = \frac{\sum_{\mathrm{supp}(\boldsymbol{k})=\boldsymbol{u}} |\hat{f}_{\boldsymbol{k}}|^2}{\sum_{\boldsymbol{k}\in\mathcal{K}\setminus\{\boldsymbol{0}\}} |\hat{f}_{\boldsymbol{k}}|^2} \in [0,1].$$

Methods

Goal

Given: N data points $\boldsymbol{x}_j \in \mathbb{T}^d$ or $[0,1]^d$ and corresponding values $y_j \in \mathbb{R}$, find a model or rather function f with

$$y_j \approx f(\boldsymbol{x}_j).$$

We can work efficiently with trigonometric models (nonuniform FFT, short: NFFT), but only in low dimensions (due to the curse of dimensionality).

ONB Approach

Ansatz: $y_j \approx f(\boldsymbol{x}_j) := \sum_{\boldsymbol{k}\in\mathcal{K}} \hat{f}_{\boldsymbol{k}}\,\mathrm{e}^{2\pi\mathrm{i}\boldsymbol{k}^\top\boldsymbol{x}_j}$, where we assume a low superposition dimension: $\boldsymbol{k}\in\mathcal{K} \iff |\mathrm{supp}(\boldsymbol{k})| \le d_s$.

1. Solve $\min_{\hat{\boldsymbol{f}}} \|\boldsymbol{y} - \Phi\hat{\boldsymbol{f}}\|_2^2 + \lambda\hat{\boldsymbol{f}}^* W \hat{\boldsymbol{f}}$ iterative (LSQR), where $\Phi = [\mathrm{e}^{2\pi\mathrm{i}\boldsymbol{k}^\top\boldsymbol{x}_j}]_{\boldsymbol{k},j}$ (fast mult. via NFFTs). $W = \mathrm{diag}(\hat{\omega}_{\boldsymbol{k}})$ (decay of $\hat{f}_{\boldsymbol{k}} \sim$ smoothness of f)
2. Compute GSI to determine active subsets $\boldsymbol{u}$. Re-compute the approximation by only keeping the active subsets. [Potts, Schmischke 2020, 2021]

Kernel-Based Approach

Search $f \in \mathcal{H} = \overline{\mathrm{span}\{\kappa(\boldsymbol{x},\cdot), \boldsymbol{x}\in\mathbb{R}^d\}}$ (RKHS).
By the representer theorem, we may solve

$$\min_{f\in\mathcal{H}} \sum_{j=1}^N (y_j - f(\boldsymbol{x}_j))^2 + \lambda\|f\|_{\mathcal{H}}^2$$

$$\iff \min_{\boldsymbol{\alpha}\in\mathbb{C}^N} \|\boldsymbol{y} - K\boldsymbol{\alpha}\|_2^2 + \lambda\boldsymbol{\alpha}^* K\boldsymbol{\alpha},$$

where $K \in \mathbb{R}^{N\times N}$ is the kernel matrix with entries $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$.
Kernel ridge regression (KRR), $y_j \approx \sum_{i=1}^N \alpha_i\,\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$.

Connection: Regularized Least Squares

Consider the case $\lambda \neq 0$ with kernels of the form $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \sum_{\boldsymbol{k}\in\mathcal{K}} \hat{\omega}_{\boldsymbol{k}}^{-1}\,\mathrm{e}^{2\pi\mathrm{i}\boldsymbol{k}^\top(\boldsymbol{x}_i-\boldsymbol{x}_j)}$.

Then, both approaches are mathematically equivalent [Shawe-Taylor, Cristianini 2004]

$$\Phi^*\Phi\hat{\boldsymbol{f}} + \lambda W\hat{\boldsymbol{f}} = \Phi^*\boldsymbol{y} \iff \hat{\boldsymbol{f}} = \lambda^{-1}W^{-1}\Phi^*(\boldsymbol{y} - \Phi\hat{\boldsymbol{f}}) = W^{-1}\Phi^*\boldsymbol{\alpha}$$

with

$$\boldsymbol{\alpha} = \lambda^{-1}(\boldsymbol{y} - \Phi\hat{\boldsymbol{f}}) \iff \lambda\boldsymbol{\alpha} = \boldsymbol{y} - \Phi W^{-1}\Phi^*\boldsymbol{\alpha} \iff (\underbrace{\Phi W^{-1}\Phi^*}_{K} + \lambda I_N)\boldsymbol{\alpha} = \boldsymbol{y}.$$

Comments and Questions

- Other orthonormal systems analogously, e.g. half-period cosine basis in non-periodic case, $\boldsymbol{x}_j \in [0,1]^d$.
- The restriction to small superposition dimensions d_s is motivated by the sparsity of effects: In practice most phenomena can be described by a few low-dim. interactions. The same applies to sufficiently smooth functions.
- In the case of $\lambda = 0$ we need to work over-/under- determined.
- In the dual setting we compute the α_j in addition. Benefit?

Results

Results - Least Squares Approach

Tested on real and synthetic data sets (regression and classification).

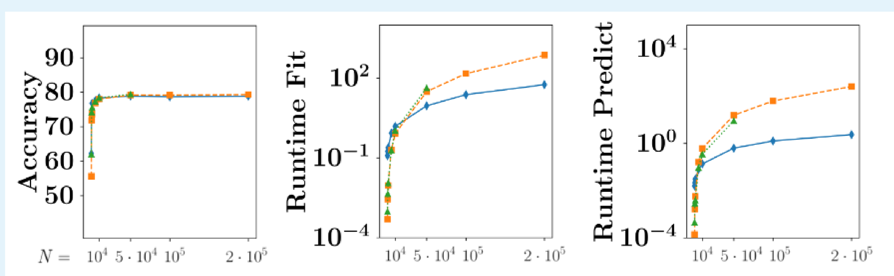
data set	d	N	error (type)	ref. method	ANOVA-LSQR
Forest Fires	12	517	12.71 (MAD)	SVM	12.65
Energy Eff. Housing	8	768	1.79 (RMSE)	Grad. Boost. Mach.	1.49
Energy Eff. Cooling	8	768	0.48 (RMSE)	Random Forest	0.44
Airfoil Self-Noise	5	1503	0.028 (rel. ℓ_2)	Sparse Rand. Features	0.016
California Housing	8	20640	0.115 (RMSE)	Local Learning Reg. NN	0.109
Ailerons	40	13750	0.0460 (RMSE)	Local Learning Reg. NN	0.0457

[Schmischke 2022, PhD thesis], publicly available julia software.

Results - Kernel Approach

We consider real data sets for binary classification and apply Gaussian ANOVA-like kernels (which we approximate by trig. polynomials).

$$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \frac{1}{n_{\text{kernels}}} \sum_{|\boldsymbol{u}|\le d_s} \mathrm{e}^{-\|\boldsymbol{x}_{i,\boldsymbol{u}} - \boldsymbol{x}_{j,\boldsymbol{u}}\|^2/\sigma^2}.$$



We compared our NFFT-based KRR with standard sklearn algorithms (KRR and SVM) in python. Example: Results for SUSY data set with $d = 18$ features. [Nestler, Stoll, Wagner 2022]

Ongoing & Future

Support Vector Regression

$$\min_{\hat{\boldsymbol{f}}} \tfrac{1}{2}\|\hat{\boldsymbol{f}}\|_2^2 + C\sum_{j=1}^N |\xi_j|^2 \quad \text{s.t. } |y_j - (\Phi\hat{\boldsymbol{f}})_j| \le \epsilon + |\xi_j|$$

Conditionally p.d. Kernels

$$f(\boldsymbol{x}) = \sum_{j=1}^N \alpha_j\kappa(\boldsymbol{x}_j,\boldsymbol{x}) + \sum_l \beta_l p_l(\boldsymbol{x}), \quad \begin{pmatrix} K & P \\ P^\top & 0 \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} \boldsymbol{y} \\ 0 \end{pmatrix}$$

Gaussian Process Regression

Use ANOVA-type kernels $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$ as covariance functions in Gaussian processes.

Refs

Software

- Julia Package ANOVAapprox by M. Schmischke:
  - `https://github.com/NFFT/ANOVAapprox.jl`
  - Bases: `"per"` (Fourier system), `"cos"` (half-per. cosine), `"cheb"` (Chebyshev polynomials), `"wav1"`...`"wav4"` (Wavelets).
  - Solvers: `"lsqr"`, `"fista"` (different regularization, group lasso), `"krr"` (solves the dual problem with CG, $K = \Phi W^{-1}\Phi^*$).
  - Interpretability: GSI and attribute ranking.
- Python code NFFT4ANOVA by T. Wagner:
  - `https://github.com/wagnertheresa/NFFT4ANOVA`
  - KRR, where the kernel is a sum of equally weighted low-dimensional Gaussian kernels.

References

1. Franziska Nestler, Martin Stoll and Theresa Wagner. Learning in High-Dimensional Feature Spaces Using ANOVA-Based Fast Matrix-Vector Multiplication. arXiv: 2111.10140, 2021. (accepted in Found. Data Sci.)
2. Daniel Potts and Michael Schmischke. Interpretable Approximation of High-Dimensional Data. SIAM J. Math. Data Sci., 3 (4), 1301–1323, 2021.
3. Daniel Potts and Michael Schmischke. Approximations of high-dimensional periodic functions with Fourier-based methods. SIAM J. Numer. Anal. 59, 2393-2429, 2021.
4. Felix Bartel, Michael Schmischke and Daniel Potts. Grouped Transformations and Regularization in High-Dimensional Explainable ANOVA Approximation. SIAM Journal on Scientific Computing, 2021 (accepted).
5. Laura Lippert, Daniel Potts and Tino Ullrich. Fast Hyperbolic Wavelet Regression meets ANOVA. ArXiv: 2108.13197, 2021.

Dr. Franziska Nestler
franziska.nestler@math.tu-chemnitz.de
http://www.tu-chemnitz.de/~nesfr

Junior Research Group SAℓE
http://www.tu-chemnitz.de/mathematik/sale

SPONSORED BY THE
Federal Ministry of Education and Research