GPy-ABCD: A Configurable Automatic Bayesian Covariance Discovery Implementation

Thomas Fletcher, Alan Bundy, Kwabena Nuamah
T.Fletcher-6@sms.ed.ac.uk, A.Bundy@ed.ac.uk, K.Nuamah@ed.ac.uk

1. Introduction

Gaussian Processes (GPs) are a very flexible class of nonparametric models which are able to fit data with very few assumptions, namely just the type of correlation (kernel) the data is expected to display.
Automatic Bayesian Covariance Discovery (ABCD)[1] is an iterative modular Gaussian Process regression framework aimed at removing the requirement for even this initial correlation form assumption.
GPy-ABCD[2] is a new implementation of an ABCD system built for ease of use and configurability; it can produce short text descriptions of fit models, it uses a revised model-space search algorithm and it removes a search bias which was required in order to retain model explainability in the original system.

2. ABCD Details

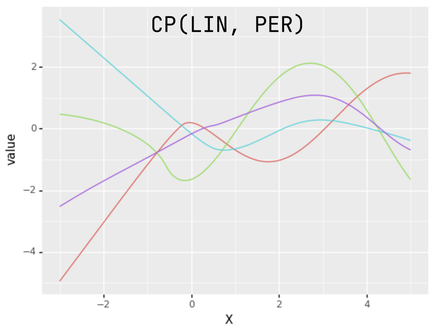
Main components:

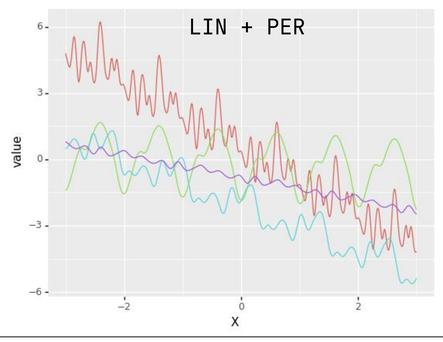
1. An open-ended and expressive language of models:
   - Base kernels: White Noise (`WN`), Constant (`C`), Linear (`LIN`), Sq. Exponential (`SE`), Periodic (`PER`)
   - Combining operators: addition, multiplication, change-point (`CP`), change-window (`CW`)
2. An efficient generation and search procedure to explore the model space:
   - A configurable beam search (a limited-bandwidth best-first-search)
   - using a context-free grammar as the successor states' generator
   - where kernel expressions self-simplify symbolically before fitting
3. A model evaluation and comparison method balancing complexity and closeness of fit:
   - BIC (Bayesian Information Criterion) by default, but arbitrary criteria allowed
4. A procedure to automatically generate descriptions of the best candidate(s):
   - Kernel expressions are rearranged to canonical sum-of-products form
   - Expression-ordering heuristics & templates generate text

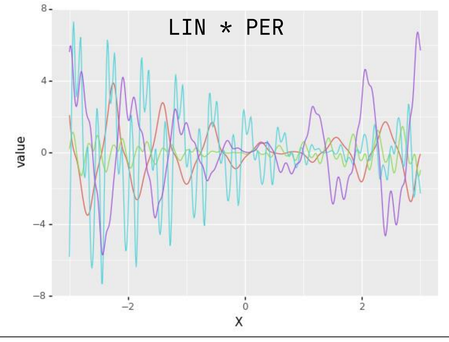
3. Kernel Expression Examples

Curves generated by Gaussian Processes with three different combinations of the same pair of base kernels:

- `CP(LIN, PER)`: Linear function transitioning to periodic function
- `LIN + PER`: Linear function with periodic component
- `LIN * PER`: Periodic function with linearly varying amplitude







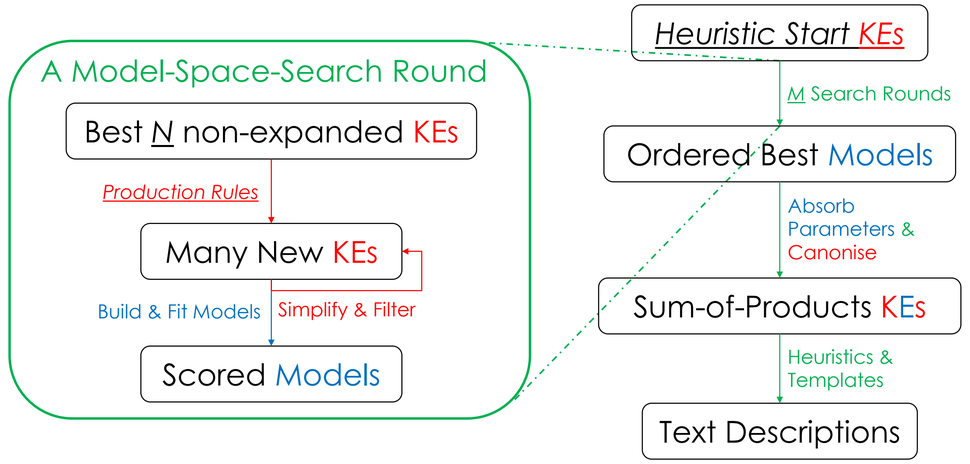
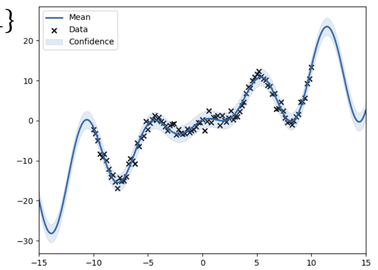


Fig.1 – GPy-ABCD's model-space search algorithm

4. Fit Example

- Synthetic data: $X = [-10{,}10]$ every 0.2, $Y = 2X\cos\left(\frac{X-5}{2}\right)^2 + \epsilon$, $\epsilon \in \{-1{,}1\}$
- Top result:

```
LIN * (PER + C)
  mul.                                |                value  |  constraints
  linear_with_offset.variance         |  0.00868862409368869  |          +ve
  linear_with_offset.offset           |  -0.08541236236743312 |
  sum.pure_std_periodic.variance      |    62.48387020446294  |          +ve
  sum.pure_std_periodic.period        |   6.2443744834482535  |          +ve
  sum.pure_std_periodic.lengthscale   |   1774.034537156076   |          +ve
  sum.bias.variance                   |   119.28065536635225  |          +ve
```



- Description:

  "The fit kernel consists of 2 components:
  - A linear function with offset -0.09; this component has overall variance 1.04
  - A periodic function with period 6.24 and lengthscale 1774.03, with linearly varying amplitude with offset -0.09; this component has overall variance 0.54"

- The runners-up are direct generalisations: `LIN * (PER + LIN)` and `LIN * LIN * (PER + C)`

5. Development Context: FRANK

GPy-ABCD was developed for the FRANK[3] (Functional Reasoning for Acquiring Novel Knowledge) query-answering system, which performs inferential and statistical reasoning on data from publicly available knowledge bases.
It uses a graph-based inference algorithm which decomposes queries into sub-queries until retrievable data is found, and then processes the nodes upwards aggregating children towards the root.
Its internal statistics expert system (SMART, i.e. Statistical Methodology Advisor at Reasoning Time) would choose to use GPy-ABCD for queries involving univariate functional shape description (e.g. "How does rainfall in the UK behave over time?" or "Is population growth in Asia periodic/linear?").

Which country will have the largest population in Africa in 2021?
MAX($y,COMP(<?x,$y>,population(?x,$y,2021): Country(?x)& location(?x,Africa)))
[Answer: Nigeria (FRANK estimate = 202550218)]

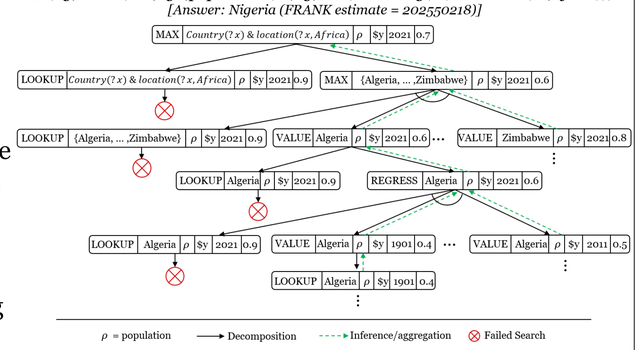


Fig.2 – Example of FRANK's inference graph

References

[1] Lloyd, James Robert; Duvenaud, David Kristjanson; Grosse, Roger Baker; Tenenbaum, Joshua B.; Ghahramani, Zoubin. "Automatic construction and natural-language description of nonparametric regression models". National Conference on Artificial Intelligence. 2014.
[2] https://github.com/T-Flet/GPy-ABCD
[3] Nuamah, Kwabena (2018): Functional inferences over heterogeneous data. PhD. University of Edinburgh.