Symbolic Regression Using Prior Knowledge

Jiří Kubalík

jiri.kubalik@cvut.cz

Symbolic Regression Using Prior Knowledge

Insufficient training data

- sparse and noisy,
- unevenly sample the input space,
- may completely omit some parts of the input space.

Models trained using only such training data tend to be

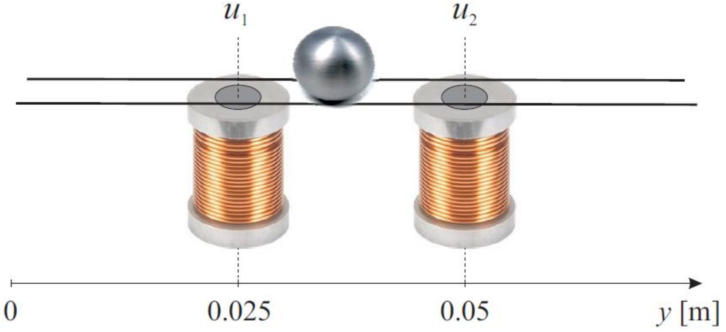
- overfitted,
- partially incorrect in terms of their steady-state characteristics or local behavior.

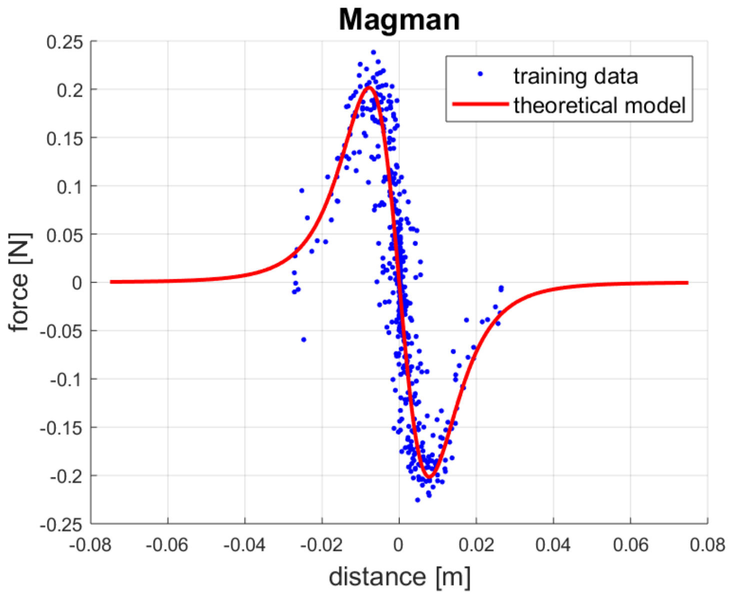
Magnetic manipulation

Magnetic manipulation – an iron ball moving along a rail and an electromagnet at a static position under the rail.

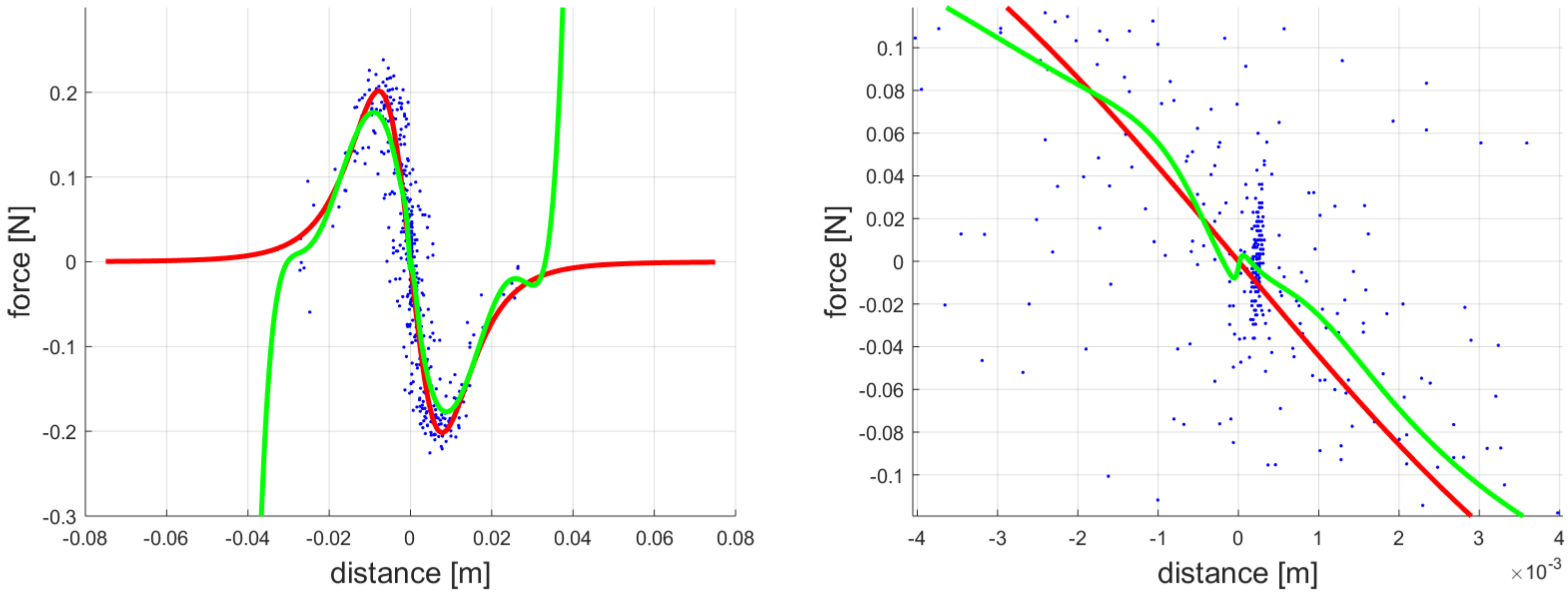
Data – noisy; only a part of the input space is covered.

Goal is to find a model of the nonlinear magnetic force affecting the ball as a function of the distance between the ball and the activated coil.





Magman: SR driven by training data only



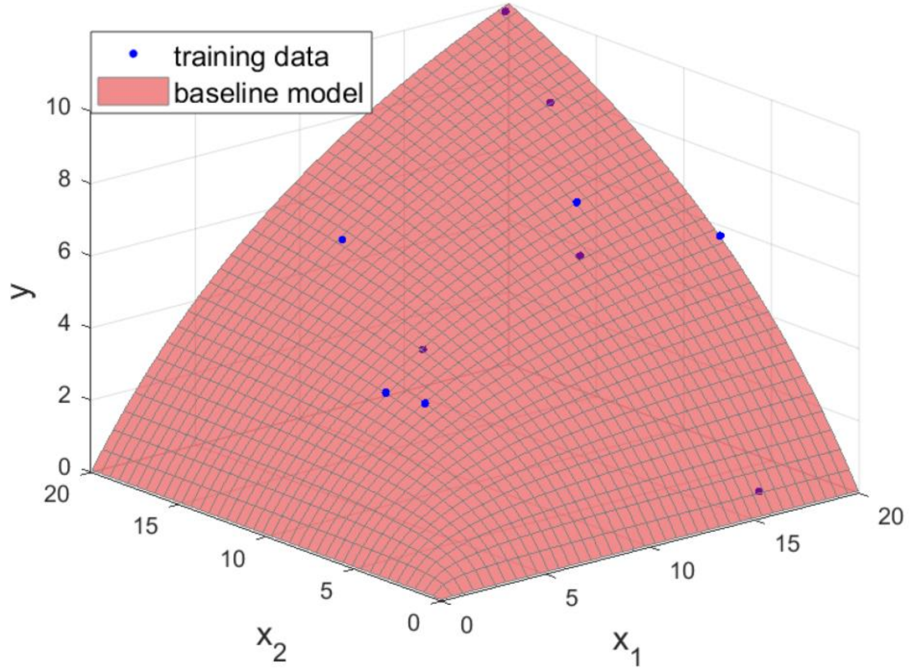
Two resistors in parallel

Resistance – equivalent resistance of two resistors in parallel.

Data – very sparse and noisy.

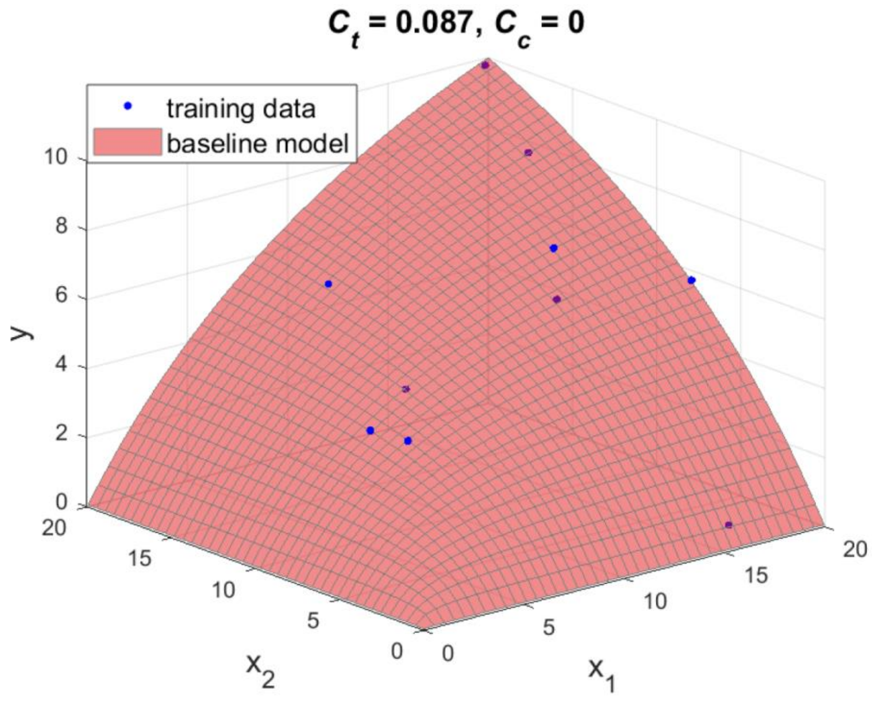
Goal is to find a model that fits the data and obeys the physical law.

Baseline model: $R = \frac{R_1 R_2}{R_1 + R_2}$

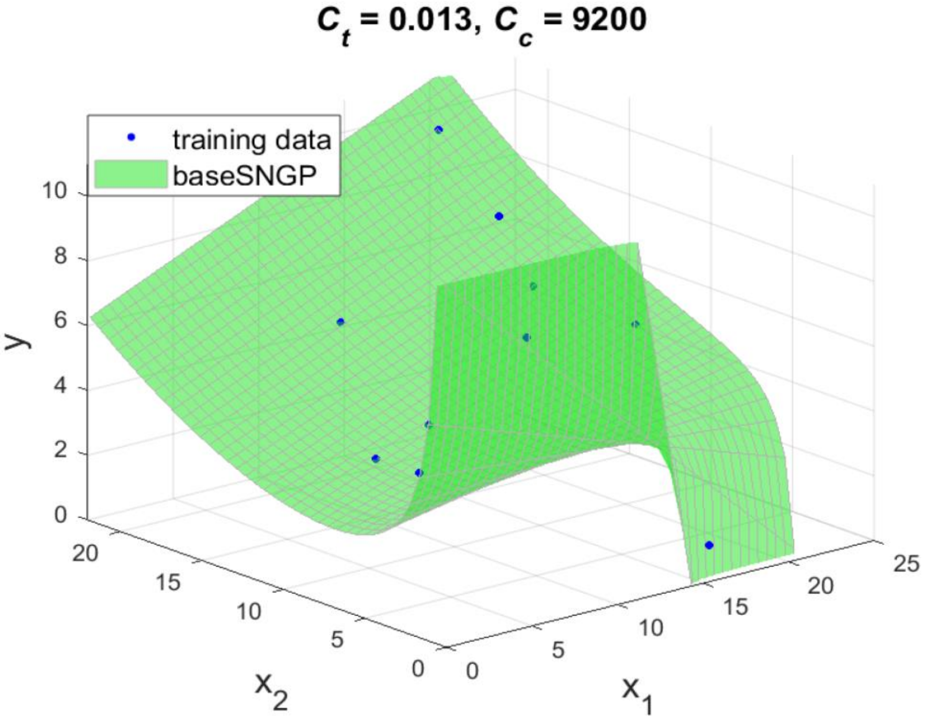


Resistance: SR driven by training data only

Baseline model

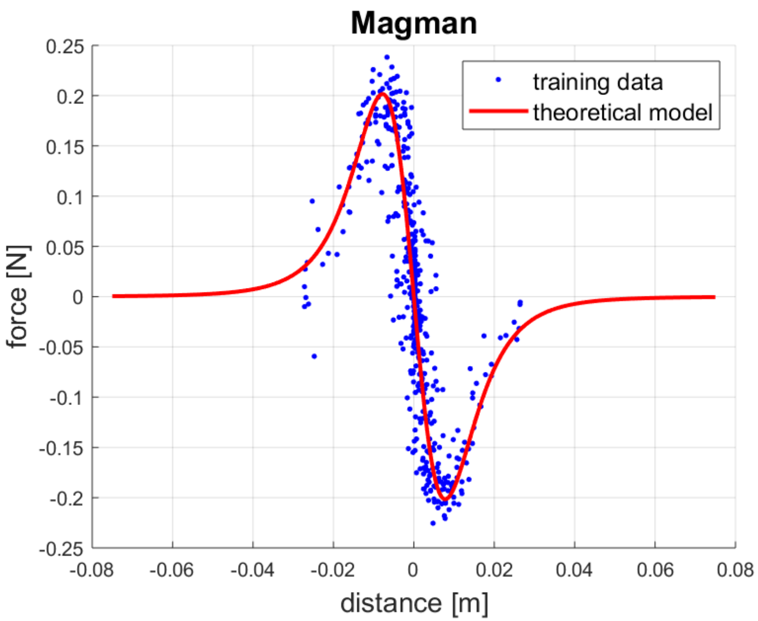


SR model



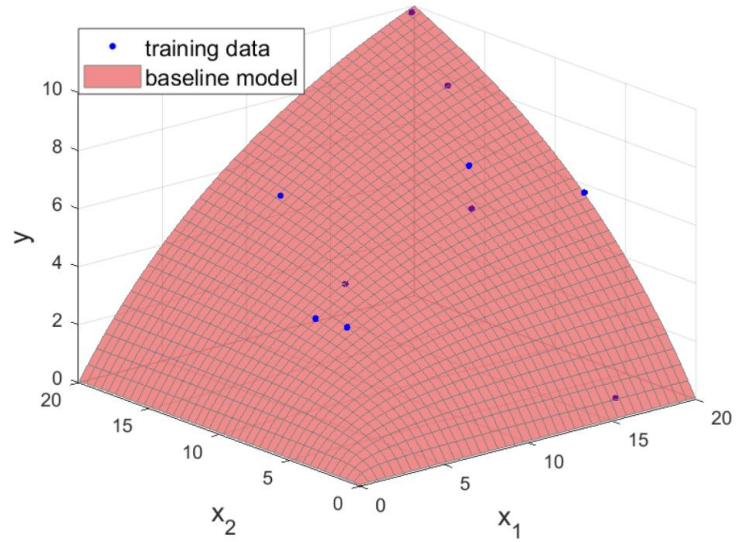
Magman: Desired model's properties

- Increasing monotonicity
  $x \in (-0.075, -0.01)$
  or
  $x \in (0.01, 0.075)$
- Decreasing monotonicity
  $x \in (-0.01, 0.01)$
- Odd symmetry
- Exact output values
  $f(-0.075) = 0.001$
  $f(0.075) = -0.001,$
  $f(0) = 0.0$



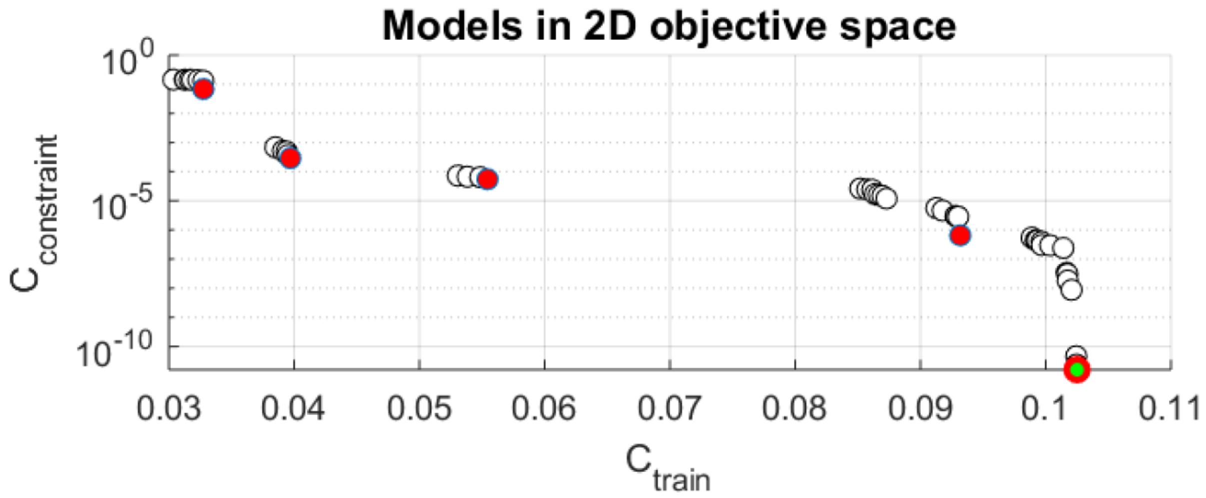
Resistance: Desired model's properties

- symmetry with respect to arguments
  $R(R_1, R_2) = R(R_2, R_1)$
- domain-specific constraint
  $R_1 = R_2 \Rightarrow R(R_1, R_2) = R_1/2$
- domain-specific constraint
  $R(R_1, R_2) \leq R_1,\ R(R_1, R_2) \leq R_2$



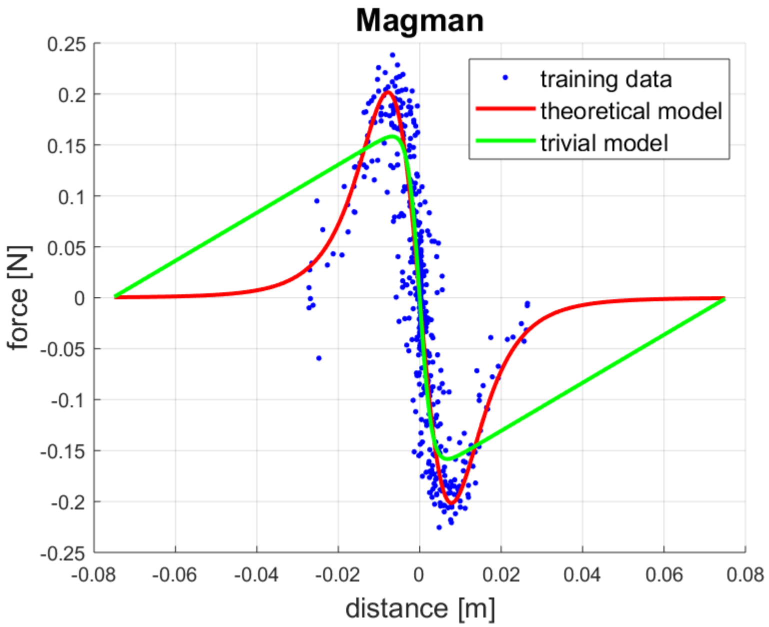
Bi-objective Symbolic Regression

- Optimisation criteria
  - minimise prediction error on training data samples
  - minimise violation of the desired model's properties
- Constraint samples set – properties are internally represented by a set of discrete data samples on which candidate models are exactly checked.
- NSGA-II – based on the concept of dominance
  - generates a set of non-dominated solutions

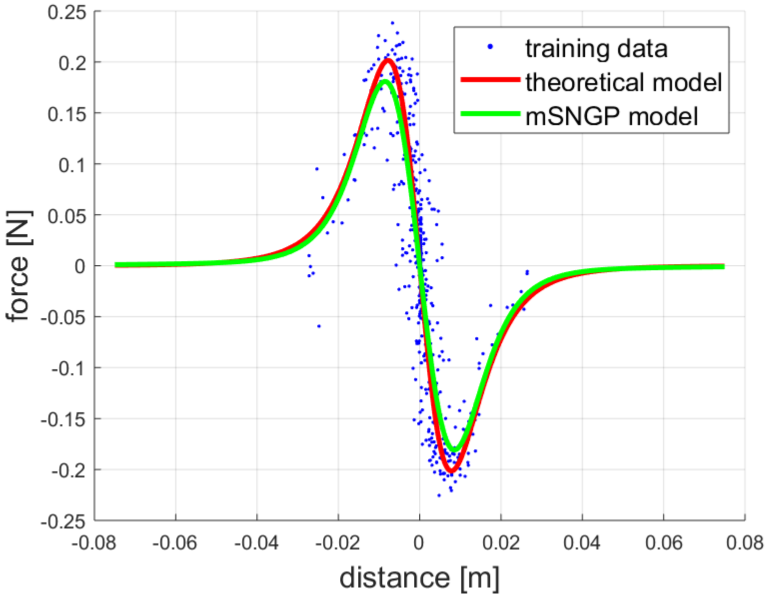


Bi-objective SR: Magman

Inaccurate, but perfectly valid

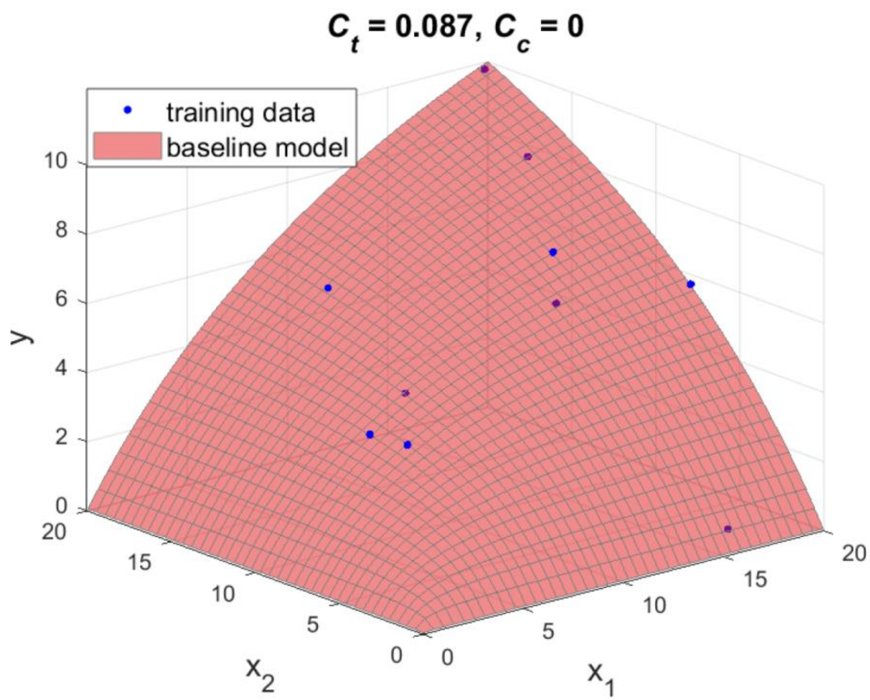


Accurate and valid

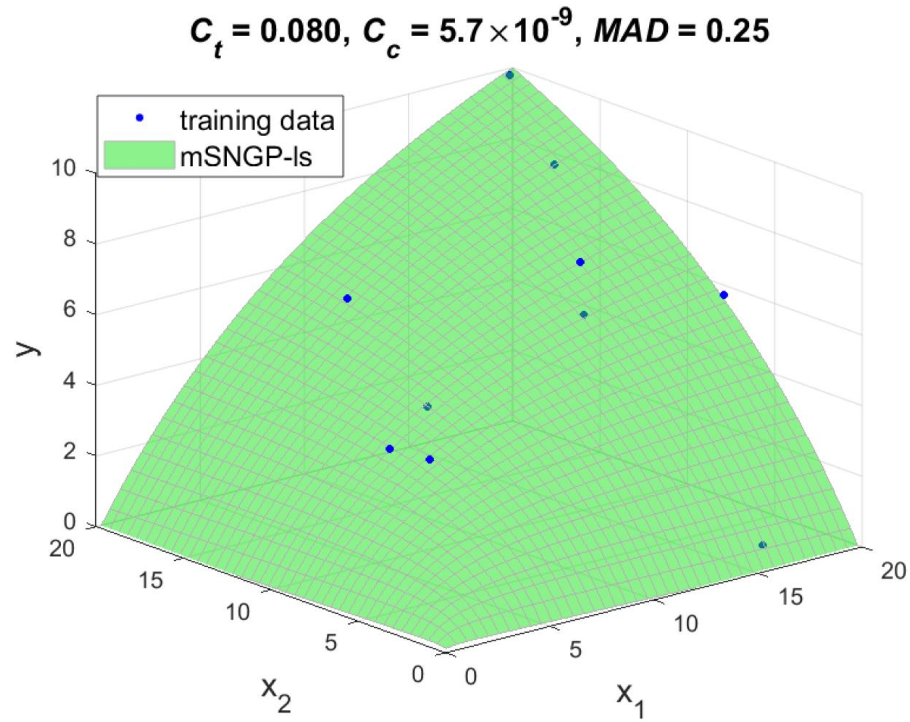


Bi-objective SR: Resistors

Baseline model



SR model



Summary

- Multi-objective SR method that produces realistic models that fit well the training data while complying with the prior knowledge of the desired model characteristics at the same time.
- Future work
    - Investigate various strategies to maintain the most relevant constraint samples during the whole run.
    - Different constraints can generate violations of a very different scale – need for some normalization.