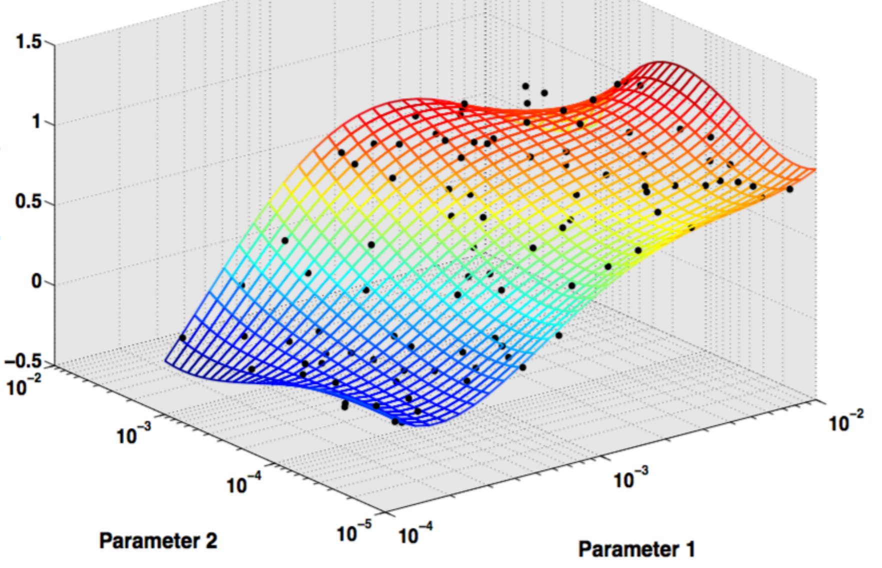
SAND2016-7019D

Polynomial surrogate construction for computational models

Sarah Teichman, Amherst College, B.A. Statistics, est. May 2018
Khachik Sargsyan, Org. 8351, Reacting Flow Research
July 27, 2016



Sandia National Laboratories

Exceptional service in the national interest

Abstract This projects investigates surrogate usage for expensive computational models. Polynomial chaos expansions are used to create inexpensive models to approximate the results of more complicated physical models. As seen through error convergence plots, error depends in part on the interaction between size of the training set, dimension of the computational model, and order of the surrogate polynomial. Generally, higher sample sizes and orders increase the accuracy of the model. However, if they get too high the model can become too complicated and over-fit the data.

Introduction

Background:

- Computational models ($\boldsymbol{f(\lambda)}$) often require many evaluations, creating the need for cheaper surrogates ($\boldsymbol{g(\lambda)}$) that approximate the results of the original model
- The goal of this project is to investigate surrogate model selection
- The current focus is on polynomial surrogate construction:

$$\mathbf{f}(\boldsymbol{\lambda}) \simeq \boldsymbol{g_c}(\boldsymbol{\lambda}) = \sum_{k=0}^{K-1} \boldsymbol{c_k \Psi_k}(\boldsymbol{\lambda})$$

- From this equation, the goal is to solve for c using least squares regression

$$\boldsymbol{c^{LSQ}} = \boldsymbol{arg_c}\min||\boldsymbol{f} - \boldsymbol{g_c}||_2$$

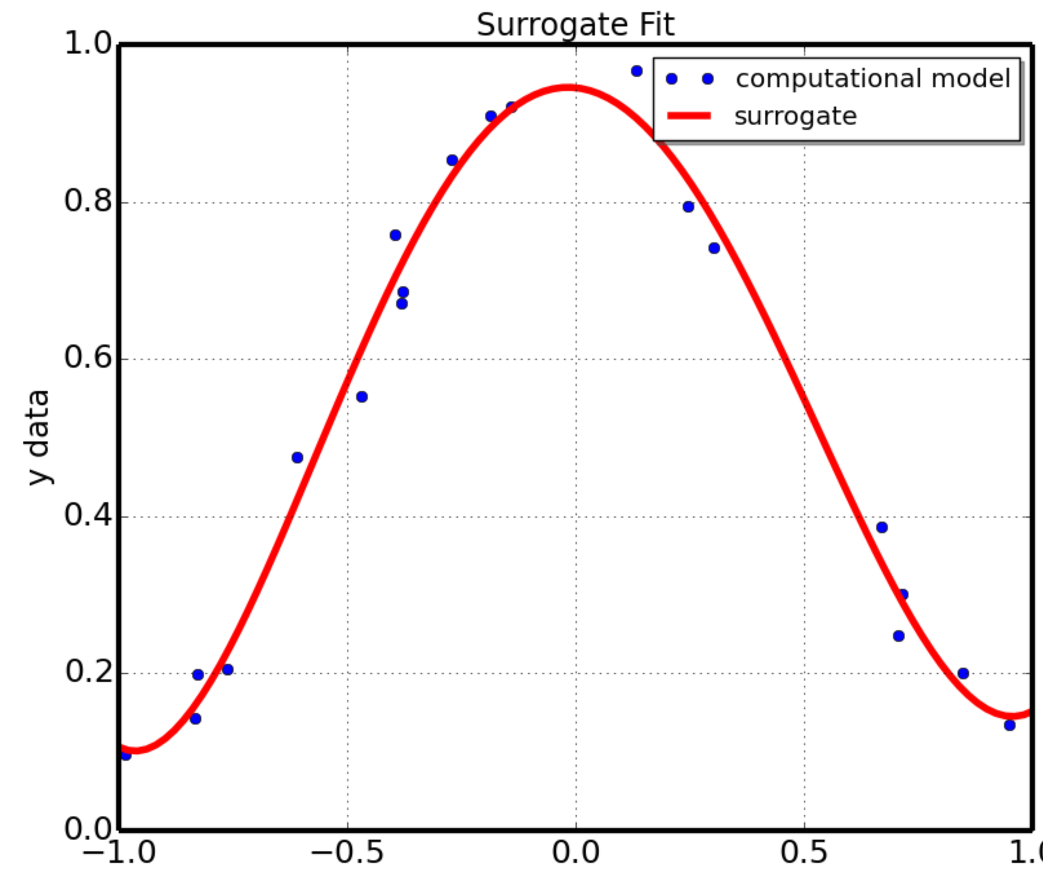
Error Computations:

- Training Error (TE): $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{f}(\boldsymbol{\lambda}^{(n)}) - \boldsymbol{g_c}(\boldsymbol{\lambda}^{(n)}))^2}$
- Validation Error (VE): same as Training Error, but with separate input values ($\lambda^{(n)}$) used for training and testing
- Leave-one-out Cross Validation (LOO): Computed by averaging the error produced by using one observation as a test and the rest as training for each observation in the dataset
- Generalized Cross-Validation (GCV): A common alternative to LOO that is less expensive to compute

Test Functions:

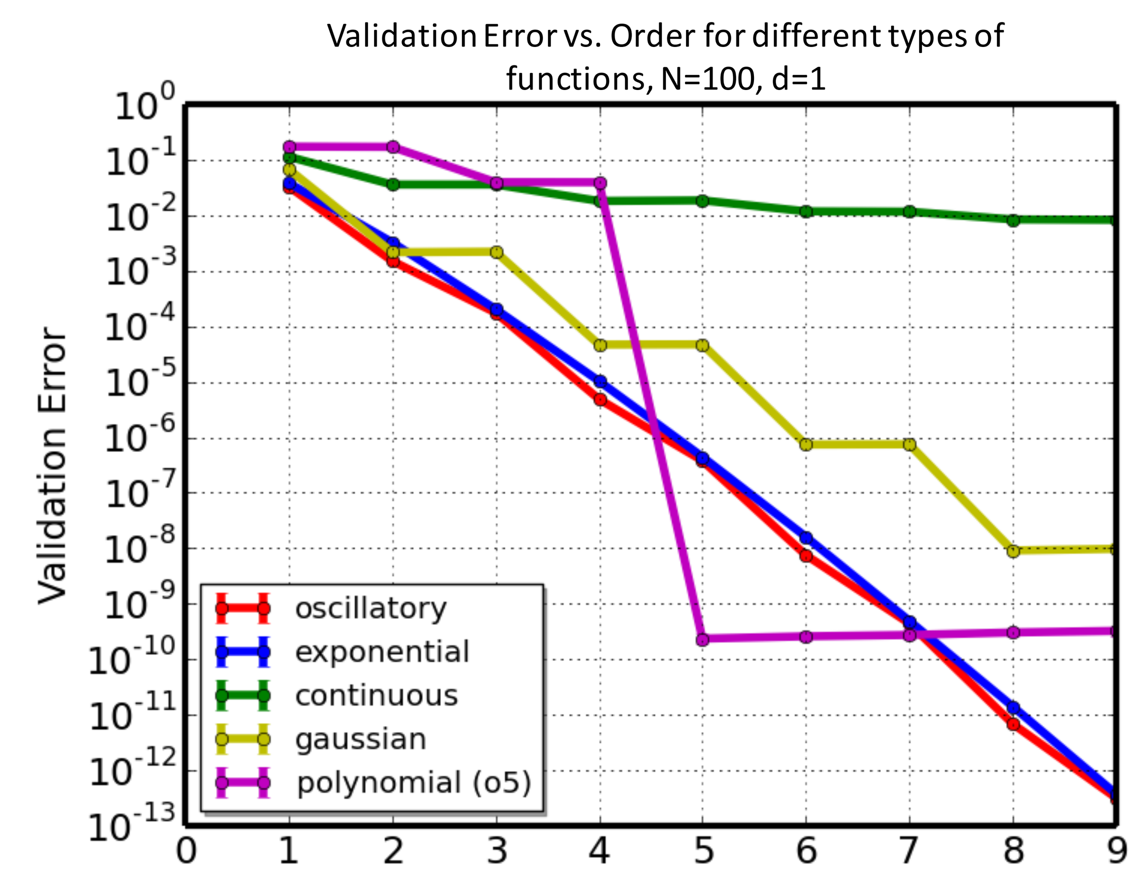
- Genz functions are a family of functions with adjustable dimensionality and smoothness
- They can be used instead of computationally heavy physical models to study surrogate construction

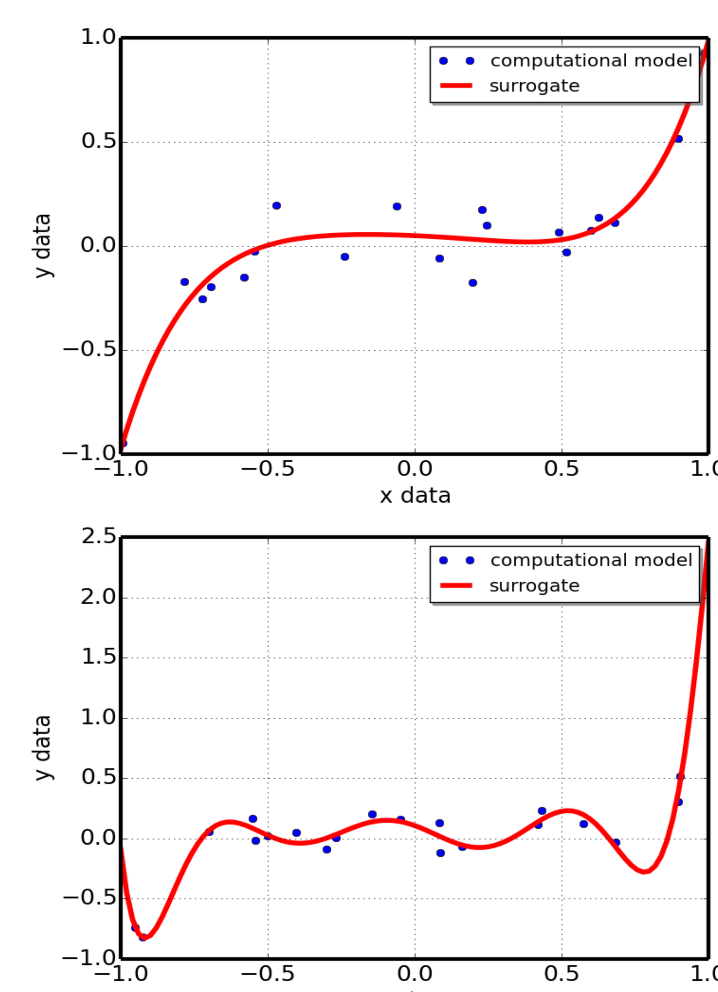
Genz Function	Formula : f(λ)
Oscillatory	$\cos(2\pi u_1 + \sum_{i=1}^{d} w_i\lambda_i)$
Exponential	$\exp(\sum_{i=1}^{d} w_i(\lambda_i - u_i))$
Continuous	$\exp(-\sum_{i=1}^{d} w_i\lvert\lambda_i - u_i\rvert)$
Gaussian	$\exp(-\sum_{i=1}^{d} w_i^2(\lambda_i - u_i)^2)$



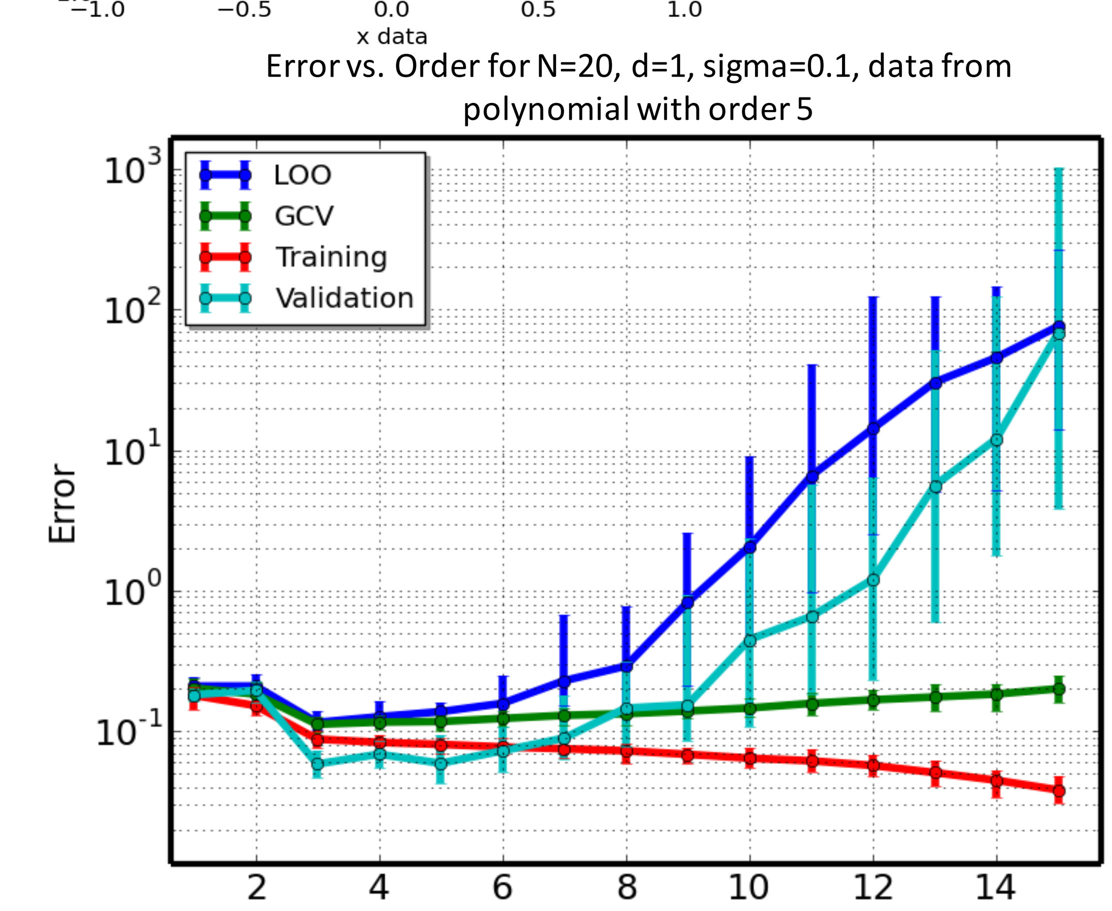
One-dimensional surrogate fit to data from the Genz-Gaussian function

Results





Overfitting Demonstration: Surrogate fits for data generated from a polynomial with the order 5 with Gaussian noise with a sigma of 0.1. The first plot uses a surrogate with the order 5, and the second plot uses a surrogate with the order 10.



Future Work

- Regularization can be implemented to reduce surrogate complexity, keeping only the most relevant terms
- Radial basis functions can be used instead of polynomials to construct surrogates
- Compressed sensing techniques can augment simple least squares regression
- Bayesian inference could be used in model selection
- Further study about why the LOO error often mimics the VE, and the GCV error mimics the TE

Acknowledgments This research project was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) through the Science Undergraduate Laboratory Internship (SULI) program.

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration