Received: 30 June 2020 | Revised: 23 March 2021 | Accepted: 25 March 2021

DOI: 10.1002/cem.3344

RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

# Genetic robust kernel sample selection for chemometric data analysis

**Fouzi Douak**[1,2] | **Noureddine Ghoggali**[2] | **Rachid Hedjam**[3] | **Mohamed Lamine Mekhalfi**[4] | **Nabil Benoudjit**[2] | **Farid Melgani**[4]

[1]Department of Industrial Engineering, University of Abbes Laghrour Khenchela, Khenchela 40004, Algeria

[2]Laboratoire d'Automatique Avancée et d'Analyse des Systèmes, Université Batna 2 Mostafa Benboulaid, Avenue Boukhlouf Med El Hadi, 05000 Batna, Algeria

[3]Department of Computer Science, Sultan Qaboos University, PO. Box 36, Alkhod 123, Muscat, Oman

[4]Department of Information Engineering and Computer Science, University of Trento, Via Sommarive 9, I-38123, Trento, Italy

**Correspondence**
Dr. Fouzi Douak, Department of Industrial Engineering, University of Abbes Laghrour Khenchela, Khenchela 40004, Algeria.
Email: douak_fouzi@univ-khenchela.dz

**Abstract**

In this work, we propose a new algorithm to improve existing techniques used in the field of spectroscopic data regression analysis. In particular, it combines the power of nonlinear kernel regressors (kernel ridge regression [KRR], kernel principal component regression [KPCR], and Gaussian process regression [GPR]) with an optimization based on nondominated sorting multi-objective genetic algorithm (NSGAII) to filter the residual outliers in the prediction space and leverage points in the features space. The proposed algorithm, contrary to most existing robust algorithms, simultaneously optimizes many complementary objectives for an automatic adaptation and thus a better outliers detection. It is well known that the elimination of outliers greatly improves the regression model. It is thus the aim of this work to develop a new robust regression algorithm. It has been applied on five different datasets, and the results are compared to both classical nonlinear regression methods and the commonly used robust regression methods robust continuum regression (RCR), partial robust M-regression (PRM), robust principal component regression (RPCR), robust PLSR (RSIMPLS), and locally weighted regression (LWR). They show that the proposed algorithm outperforms the classical nonlinear regression methods and is a promising competitor to the robust methods outperforming most of them. Even though the results obtained are only from five datasets, this algorithm can be considered an interesting contribution for improving data analysis in the field of chemometrics.

**KEYWORDS**

Gaussian process regression (GPR), kernel principal component regression (KPCR), kernel ridge regression (KRR), outliers, sample selection

## 1 | INTRODUCTION

Spectroscopy has shown great importance in product analysis and quality control of complex chemical and physical systems, thanks to its simplicity, speed, nondestructive analytical nature, and multivariate calibration methods. Thus, spectroscopic data can be opportunely availed to draw quantitative and qualitative information, which is doable via regression and classification models.[1–3]

  