CRC FOCUS

# SWARM INTELLIGENCE METHODS FOR STATISTICAL REGRESSION

Soumya D. Mohanty

CRC Press
Taylor & Francis Group

# Swarm Intelligence Methods for Statistical Regression

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

# Swarm Intelligence Methods for Statistical Regression

Soumya D. Mohanty


CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742



Printed on acid-free paper
Version Date: 20181128

International Standard Book Number-13: 978-1-138-55818 (Hardback)



**Visit the Taylor & Francis Web site at**
**http://www.taylorandfrancis.com**

**and the CRC Press Web site at**
**http://www.crcpress.com**

*Dedication*
*To my parents, Rama Ranjan and Krishna*

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

# Contents

# Preface

This book is based on a set of lectures on big data analysis delivered at the BigDat International Winter School held at Bari, Italy, in 2017. The lectures focused on a very practical issue encountered in the statistical regression of non-linear models, namely, the numerical optimization of the fitting function. The optimization problem in statistical analysis, especially in big data applications, is often a bottleneck that forces either the adoption of simpler models or a shift to linear models even where non-linearity is known to be a better option. The goal of the lectures was to introduce the audience to a set of relatively recent biology inspired stochastic optimization methods, collectively called swarm intelligence (SI) methods, that are proving quite effective in tackling the optimization challenge in statistical analysis.

It was clear from the audience response at these lectures that, despite their collective background in very diverse areas of data analysis ranging from the natural sciences to the social media industry, many had not heard of and none had seriously explored SI methods. The root causes behind this lacuna seem to be (a) a lack of familiarity, within the data analysis community, of the latest literature in stochastic optimization, and (b) lack of experience and guidance in tuning these methods to work well in real-world problems. Matters are not helped by the fact that there are not a whole lot of papers in the optimization community examining the role of SI methods in statistical analysis, most of the focus in that field being on optimization problems in engineering.

I hope this small book helps in bridging the current divide between the two communities. As I have seen within my own

research, the statistical data analyst will find that success in solving the optimization challenge spurs the contemplation of better, more sophisticated models for data. Students and researchers in the SI community reading this book will find that statistical data analysis offers a rich source of challenging test beds for their methods.

The aim of the book is to arm the reader with practical tips and rules of thumb that are observed to work well, not to provide a deeper than necessary theoretical background. In particular, the book does not delve deep into the huge range of SI methods out there, concentrating rather on one particular method, namely particle swarm optimization (PSO). My experience in teaching SI methods to students has shown that it is best to learn these methods by starting with one and understanding it well. For this purpose, PSO provides the simplest entry point. Similarly, this book does not provide a more than superficial background in optimization theory, choosing to only highlight some its important results.

It is assumed that the reader of this book has a basic background in probability theory and function approximation at the level of undergraduate or graduate courses. Nonetheless, appendices are provided that cover the required material succinctly. Instead of a problem set, two realistic statistical regression problems covering both parametric and nonparametric approaches form the workhorse exercises in this book. The reader is highly encouraged to independently implement these examples and reproduce the associated results provided in the book.

References are primarily provided in the "Notes" section at the end of each chapter. While it was tempting to include in the book a more complete review of the technical literature than what is currently provided, I decided to point the reader to mostly textbooks or review articles. This was done keeping in mind the expected readership of this book, which I assume would be similar to the student-heavy makeup of the BigDat participants. This inevitably means that many key references have been left out but I hope that the ones included will give

readers a good start in seeking out the technical material that is appropriate for their application areas.

*Acknowledgements*: It is a pleasure to acknowledge colleagues who supported the inception and development of this book. I thank Carlos Martin-Vide, Donato Malerba, and other organizers of the BigDat schools for inviting me as a lecturer and for their hospitality. This annual school provides a wonderful and refreshing opportunity to interact with data scientists and students from a broad spectrum of fields, and I hope that it continues to have several future editions.

Before embarking on the writing of this book, I had the pleasure of discussing the core content of the lectures with Innocenzo Pinto and Luigi Troiano at the University of Sannio at Benevento, Italy. I had similar opportunities, thanks to Runqiu Liu and Zong-kuan Guo at the Chinese Academy of Sciences, Beijing, to present the material to undergraduate and graduate students during my lectures there on gravitational wave data analysis. I am greatly indebted to Yan Wang at Huazhong University of Science and Technology, Wuhan, for a thorough reading of the draft and many invaluable comments and suggestions. Several useful comments from Ram Valluri at the University of Western Ontario are appreciated as well. Finally, I thank Randi Cohen at CRC press for contacting me and initiating this project.